Upre Vinod, DS20SEP03, prabhasbablu123@gmail.com

In [1]:
```
import numpy as np
import pandas as pd
```

In [156]:
```
df = pd.read_csv("D:/New folder/GooglePlaystore.csv")
```

In [157]:
```
df.head()
```

Out[157]:

	App	Category	Rating	Reviews	Size	Installs	Type	Price	Content Rating	Genres	Last Updated	Current Ver	Android Ver
0	Photo Editor & Candy Camera & Grid & ScrapBook	ART_AND_DESIGN	4.1	159	19000.0	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	14000.0	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	8700.0	5,000,000+	Free	0	Everyone	Art & Design	August 1, 2018	1.2.4	4.0.3 and up
3	Sketch - Draw & Paint	ART_AND_DESIGN	4.5	215644	25000.0	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	2800.0	100,000+	Free	0	Everyone	Art & Design;Creativity	June 20, 2018	1.1	4.4 and up

Inbox (1,413) - prabhasbablu123 | Downloads/ | Python_Assignment_1 - Jupyter

localhost:8888/notebooks/Downloads/Python_Assignment_1.ipynb

jupyter Python_Assignment_1 Last Checkpoint: 01/29/2021 (autosaved)

Logout

File Edit View Insert Cell Kernel Widgets Help

Trusted | Python 3

Run | Markdown

In [158]:
```
df.shape
```

Out[158]:
```
(10841, 13)
```

In [159]:
```
df.dtypes
```

Out[159]:
```
App               object
Category          object
Rating           float64
Reviews           object
Size             float64
Installs          object
Type              object
Price             object
Content Rating    object
Genres            object
Last Updated      object
Current Ver       object
Android Ver       object
dtype: object
```

In [160]:
```
df.columns
```

Out[160]:
```
Index(['App', 'Category', 'Rating', 'Reviews', 'Size', 'Installs', 'Type',
       'Price', 'Content Rating', 'Genres', 'Last Updated', 'Current Ver',
       'Android Ver'],
      dtype='object')
```

In [161]:
```
df.describe()
```

Type here to search

ENG INTL 8:49 PM 15-Feb-21

In [161]:
```
df.describe()
```

Out[161]:

	Rating	Size
count	9367.000000	10841.000000
mean	4.193338	21516.529524
std	0.537431	20746.537567
min	1.000000	8.500000
25%	4.000000	5900.000000
50%	4.300000	18000.000000
75%	4.500000	26000.000000
max	19.000000	100000.000000

Data clean up – Missing value treatment

In [162]:
```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 10841 entries, 0 to 10840
Data columns (total 13 columns):
 #   Column          Non-Null Count  Dtype
---  ------          --------------  -----
 0   App             10841 non-null  object
 1   Category        10841 non-null  object
 2   Rating          9367 non-null   float64
 3   Reviews         10841 non-null  object
```

jupyter Python_Assignment_1 Last Checkpoint: 01/29/2021 (autosaved) Logout

File Edit View Insert Cell Kernel Widgets Help Trusted Python 3

Run Markdown

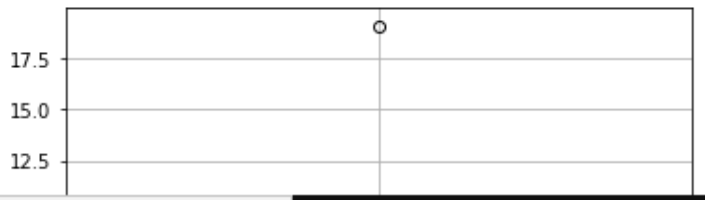
```
dtypes: float64(2), object(11)
memory usage: 1.1+ MB
```

In [163]:
```
df.isnull().sum()
```

Out[163]:
```
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 [164]:
```
import matplotlib.pyplot as plt
%matplotlib inline
df.boxplot('Rating')
```

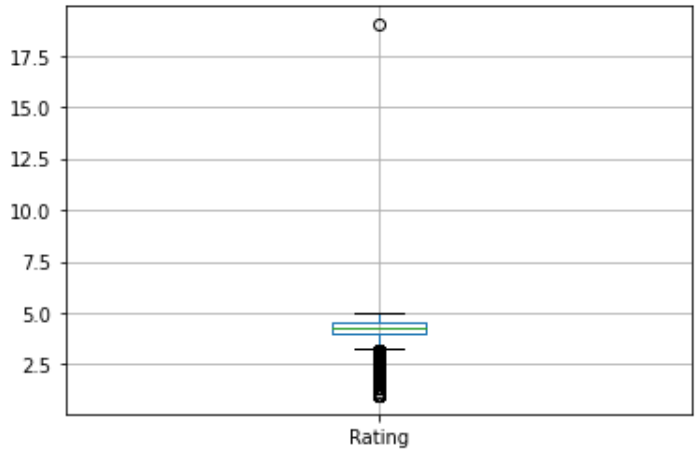
Out[164]:
```
<matplotlib.axes._subplots.AxesSubplot at 0x23675ba2bb0>
```



```
dtype: int64
```

In [164]:
```
import matplotlib.pyplot as plt
%matplotlib inline
df.boxplot('Rating')
```

Out[164]:
```
<matplotlib.axes._subplots.AxesSubplot at 0x23675ba2bb0>
```



a. Drop records where rating is missing since rating is our target/study variable

In [165]:
```
df[df.Rating>5]
```

Out[165]:

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

a. Drop records where rating is missing since rating is our target/study variable

In [165]:
```
df[df.Rating>5]
```

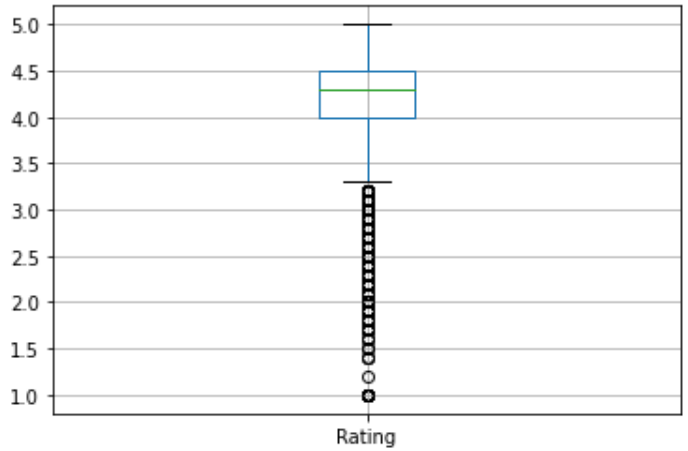
Out[165]:

	App	Category	Rating	Reviews	Size	Installs	Type	Price	Content Rating	Genres	Last Updated	Current Ver	Android Ver
10472	Life Made WI-Fi Touchscreen Photo Frame	1.9	19.0	3.0M	21516.529524	Free	0	Everyone	NaN	February 11, 2018	1.0.19	4.0 and up	NaN

In [166]:
```
df.drop([10472],inplace=True)
```

In [167]:
```
df.boxplot('Rating')
```

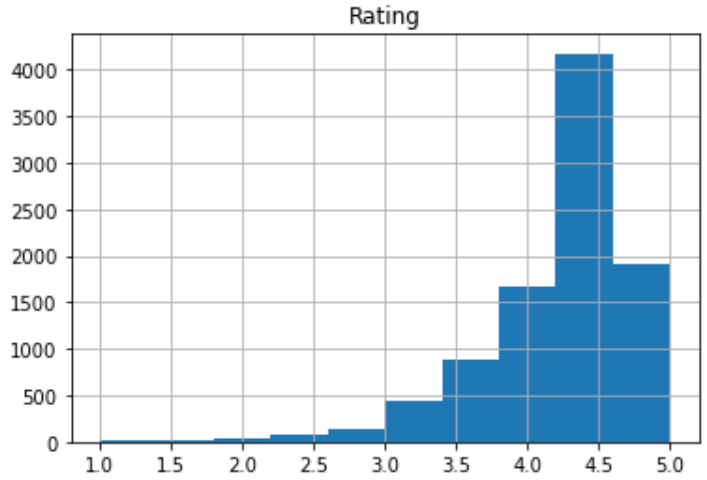
Out[167]:
```
<matplotlib.axes._subplots.AxesSubplot at 0x236760eb1c0>
```



Rating

In [168]:
```
df.hist('Rating')
```

Out[168]:
```
array([[<matplotlib.axes._subplots.AxesSubplot object at 0x0000023676141D90>]],
      dtype=object)
```



Type *Markdown* and LaTeX: α^2

In [171]:
```
#Drop records where rating is missing since rating is our target/study variable
df.isnull().sum()
```

Out[171]:
```
App             0
Category        0
```

In [171]:
```
#Drop records where rating is missing since rating is our target/study variable
df.isnull().sum()
```

Out[171]:
```
App                0
Category           0
Rating          1474
Reviews            0
Size               0
Installs           0
Type               1
Price              0
Content Rating     0
Genres             0
Last Updated       0
Current Ver        8
Android Ver        2
dtype: int64
```

In [172]:
```
df.dropna(axis=0,subset=['Rating'],inplace=True)
```

b. Check the null values for the Android Ver column. i. Are all 3 records having the same problem? ii. Drop the 3rd record i.e. record for “Life Made WIFI …” iii. Replace remaining missing values with the mode

c. Current ver – replace with most common value

In [173]:
```
df['Current Ver'].fillna(str(df['Current Ver'].mode().values[0]), inplace=True)
df['Android Ver'].fillna(str(df['Android Ver'].mode().values[0]), inplace=True)
```

correcting the datatypes

In [175]:
```
df.head()
```

Out[175]:

	App	Category	Rating	Reviews	Size	Installs	Type	Price	Content Rating	Genres	Last Updated	Current Ver	Android Ver
0	Photo Editor & Candy Camera & Grid & ScrapBook	ART_AND_DESIGN	4.1	159	19000.0	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	14000.0	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	8700.0	5,000,000+	Free	0	Everyone	Art & Design	August 1, 2018	1.2.4	4.0.3 and up
3	Sketch - Draw & Paint	ART_AND_DESIGN	4.5	215644	25000.0	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	2800.0	100,000+	Free	0	Everyone	Art & Design;Creativity	June 20, 2018	1.1	4.4 and up

In [176]:
```
df.dtypes
```

Out[176]:
```
App               object
Category          object
Rating           float64
Reviews           object
Size             float64
Installs          object
Type              object
```

jupyter Python_Assignment_1 Last Checkpoint: 01/29/2021 (autosaved) Logout

File Edit View Insert Cell Kernel Widgets Help Trusted Python 3

Run Markdown

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

In [177]:
```
#price,installs,Reviews these are the columns we have to conver in numeric
df['Price'] = df['Price'].apply(lambda x: str(x).replace('$', '') if '$' in str(x) else str(x))
df['Price'] = df['Price'].apply(lambda x: float(x))
```

In [178]:
```
df['Installs'] = df['Installs'].apply(lambda x: str(x).replace('+', '') if '+' in str(x) else str(x))
df['Installs'] = df['Installs'].apply(lambda x: str(x).replace(',', '') if ',' in str(x) else str(x))
df['Installs'] = df['Installs'].apply(lambda x: float(x))
```

In [179]:
```
df['Reviews'] = pd.to_numeric(df['Reviews'], errors='coerce')
```

In [180]:
```
df.dtypes
```

Out[180]:
```
App               object
Category          object
Rating           float64
Reviews            int64
Size             float64
Installs         float64
```

Inbox (1,413) - prabhasbablu123 | Downloads/ | Python_Assignment_1 - Jupyter

localhost:8888/notebooks/Downloads/Python_Assignment_1.ipynb

jupyter Python_Assignment_1 Last Checkpoint: 01/29/2021 (autosaved)

Logout

File Edit View Insert Cell Kernel Widgets Help

Trusted | Python 3

Run | Markdown

In [181]:
```
df['Rating'].value_counts()
```

Out[181]:

4.4	1109
4.3	1076
4.5	1038
4.2	952
4.6	823
4.1	708
4.0	568
4.7	499
3.9	386
3.8	303
5.0	274
3.7	239
4.8	234
3.6	174
3.5	163
3.4	128
3.3	102
4.9	87
3.0	83
3.1	69
3.2	64
2.9	45
2.8	42
2.6	25
2.7	25
2.5	21
2.3	20
2.4	19
1.0	16

Type here to search

ENG INTL 8:50 PM 15-Feb-21

In [182]:
```
df['Installs'].value_counts()
```

Out[182]:
```
1.000000e+06    1577
1.000000e+07    1252
1.000000e+05    1150
1.000000e+04    1010
5.000000e+06     752
1.000000e+03     713
5.000000e+05     538
5.000000e+04     467
5.000000e+03     432
1.000000e+08     409
1.000000e+02     309
5.000000e+07     289
5.000000e+02     201
5.000000e+08      72
1.000000e+01      69
1.000000e+09      58
5.000000e+01      56
5.000000e+00       9
1.000000e+00       3
Name: Installs, dtype: int64
```

In [183]:
```
df[df.Reviews>df.Installs].index
```

Out[183]:
```
Int64Index([2454, 4663, 5917, 6700, 7402, 8591, 10697], dtype='int64')
```

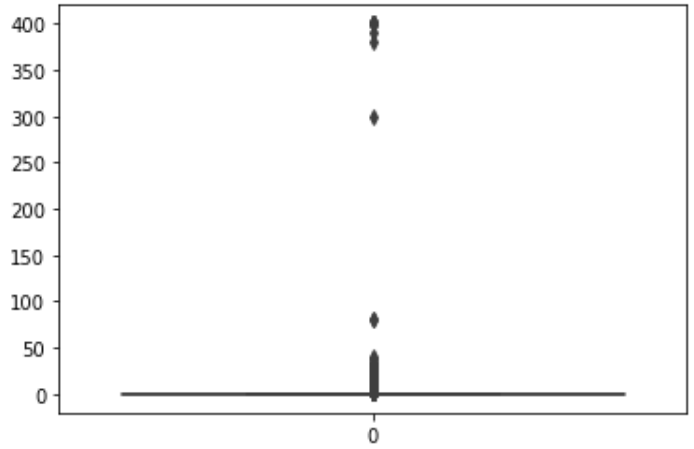
In [184]:
```
df.drop([2454, 4663, 5917, 6700, 7402, 8591, 10697],inplace=True)
```

Identify and handle outliers

Identify and handle outliers

In [185]:
```
#price column
import seaborn as sns
sns.boxplot(data=df['Price'])
```

Out[185]: <matplotlib.axes._subplots.AxesSubplot at 0x236761c7d00>



In [186]:
```
df.drop(df[df.Price>200].index,axis=0,inplace=True)
```

In [187]:
```
df.shape
```

Out[187]: (9344, 13)

In [188]:
```
df[df.Price>30].index
```

Inbox (1,413) - prabhasbablu123 | Downloads/ | Python_Assignment_1 - Jupyter

localhost:8888/notebooks/Downloads/Python_Assignment_1.ipynb

jupyter Python_Assignment_1 Last Checkpoint: 01/29/2021 (autosaved)

Logout

File Edit View Insert Cell Kernel Widgets Help

Trusted Python 3

Run Markdown

In [186]:
```
df.drop(df[df.Price>200].index,axis=0,inplace=True)
```

In [187]:
```
df.shape
```

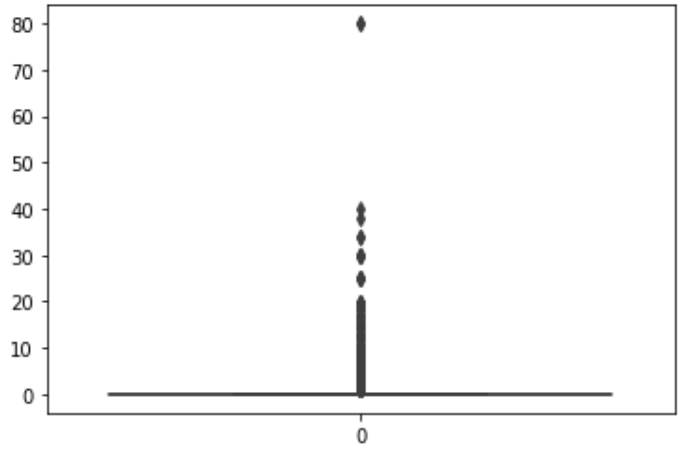
Out[187]: (9344, 13)

In [188]:
```
df[df.Price>30].index
```

Out[188]: Int64Index([2253, 2301, 2365, 2402, 2414, 5360], dtype='int64')

In [189]:
```
sns.boxplot(data=df['Price'])
```

Out[189]: <matplotlib.axes._subplots.AxesSubplot at 0x236762e4f40>



In [190]: df.drop(df[df.Price>30].index,axis=0,inplace=True)

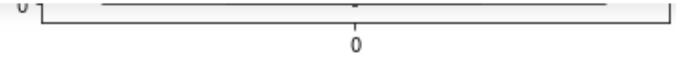
Type here to search

ENG INTL 8:51 PM 15-Feb-21

jupyter Python_Assignment_1 Last Checkpoint: 01/29/2021 (autosaved) Logout

File Edit View Insert Cell Kernel Widgets Help Trusted Python 3

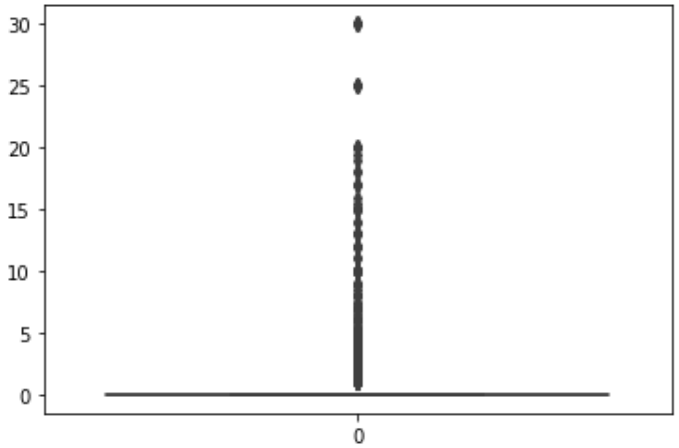
Run Markdown



In [190]:
```
df.drop(df[df.Price>30].index,axis=0,inplace=True)
```

In [191]:
```
sns.boxplot(data=df['Price'])
```

Out[191]:
```
<matplotlib.axes._subplots.AxesSubplot at 0x236761c3eb0>
```



In [192]:
```
sns.distplot(df['Price'])
```

```
C:\Users\vikas\anaconda3\lib\site-packages\seaborn\distributions.py:369: UserWarning: Default bandwidth for data is 0; skipping density estimation.
  warnings.warn(msg, UserWarning)
```

Out[192]:
```
<matplotlib.axes._subplots.AxesSubplot at 0x236762a9460>
```

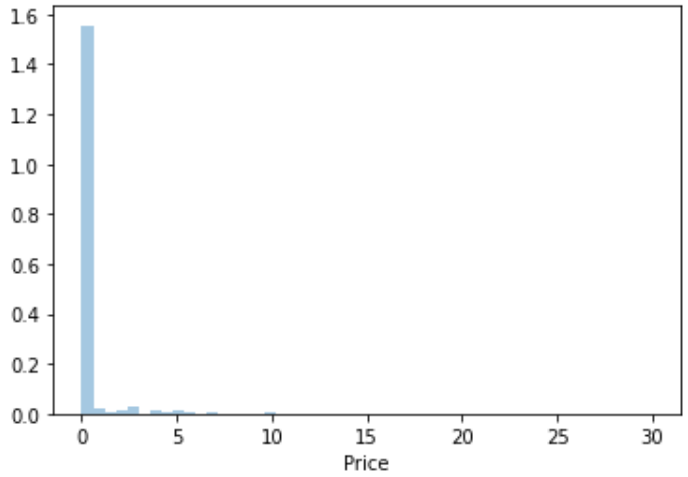
jupyter Python_Assignment_1 Last Checkpoint: 01/29/2021 (autosaved) Logout

File Edit View Insert Cell Kernel Widgets Help Trusted Python 3

Run Markdown

```
density estimation.
  warnings.warn(msg, UserWarning)
```

Out[192]: <matplotlib.axes._subplots.AxesSubplot at 0x236762a9460>



In [193]:
```
##Reviews column
sns.boxplot(data=df['Reviews'])
```

Out[193]: <matplotlib.axes._subplots.AxesSubplot at 0x23676315970>

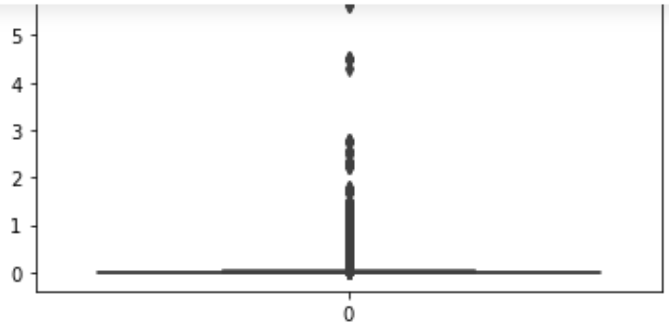
jupyter Python_Assignment_1 Last Checkpoint: 01/29/2021 (autosaved)

Logout

File Edit View Insert Cell Kernel Widgets Help

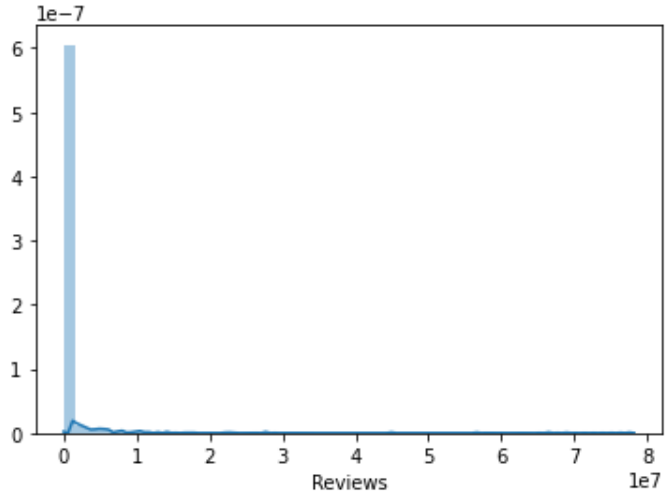
Trusted Python 3

Run Markdown



In [194]:
```
sns.distplot(df['Reviews'])
```

Out[194]: <matplotlib.axes._subplots.AxesSubplot at 0x23676428610>



Inbox (1,413) - prabhasbablu123 | Downloads/ | Python_Assignment_1 - Jupyter

localhost:8888/notebooks/Downloads/Python_Assignment_1.ipynb

jupyter Python_Assignment_1 Last Checkpoint: 01/29/2021 (autosaved)

Logout

File Edit View Insert Cell Kernel Widgets Help

Trusted | Python 3

Run | Markdown

Reviews 167

In [195]:
```
df1=df[df.Reviews<1000000]
```

In [196]:
```
df1.shape
```

Out[196]:
```
(8634, 13)
```

In [197]:
```
##Installs column
df['Installs'].describe()
```

Out[197]:
```
count    9.338000e+03
mean     1.795108e+07
std      9.136965e+07
min      5.000000e+00
25%      1.000000e+04
50%      5.000000e+05
75%      5.000000e+06
max      1.000000e+09
Name: Installs, dtype: float64
```

In [198]:
```
np.percentile(df['Installs'],95)
```

Out[198]:
```
100000000.0
```

In [200]:
```
for i in range(0,101,1):
    print(' The {} percentile of installs is {} '.format(i,np.percentile(df['Installs'],i)))
```

```
 The 0 percentile of installs is 5.0
 The 1 percentile of installs is 50.0
 The 2 percentile of installs is 100.0
```

Type here to search

ENG INTL 8:51 PM 15-Feb-21

Inbox (1,413) - prabhasbablu123 | Downloads/ | Python_Assignment_1 - Jupyter

localhost:8888/notebooks/Downloads/Python_Assignment_1.ipynb

jupyter Python_Assignment_1 Last Checkpoint: 01/29/2021 (autosaved)

Logout

File Edit View Insert Cell Kernel Widgets Help

Trusted | Python 3

Run | Markdown

In [201]:
```
df[df.Installs>np.percentile(df['Installs'],95)].index
```

Out[201]:
```
Int64Index([ 152,  335,  336,  338,  340,  341,  342,  347,  371,  378,
             ...
            4222, 4234, 4365, 4566, 4676, 5395, 5596, 5856, 7536, 9844],
           dtype='int64', length=130)
```

In [202]:
```
df.drop(df[df.Installs>np.percentile(df['Installs'],95)].index,axis=0,inplace=True)
```

In [203]:
```
df.shape
```

Out[203]:
```
(9208, 13)
```

In [ ]:

Data Analysis to answer business questions

In [206]:
```
#5. What is the distribution of ratings like? (use Seaborn) More skewed towards higher/lower values?
#a. How do you explain this? ,b. What is the implication of this on your analysis?

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

Out[206]:
```
<matplotlib.axes._subplots.AxesSubplot at 0x23676504280>
```

Type here to search

ENG INTL 8:51 PM 15-Feb-21

```
print("The skewnes of this data distribution is",df['Rating'].skew())
```

```
The skewnes of this data distribution is -1.8425872652320243
```

In [208]:
```
#the data left skewed so the median of this distribution is > mean of the distribution
print('The Median of this distribution {} is greater than mean {} of this distribution'.
      format(df.Rating.median(),df.Rating.mean()))
```

```
The Median of this distribution 4.3 is greater than mean 4.190117289313644 of this distribution
```

In [ ]:
```
#so the skewness between -1 to 1 it states data is highly skewed towards lower values
```

In [210]:
```
#6. What are the top Content Rating values?
  #a. Are there any values with very few records?
  #b. If yes, drop those as they won't help in the analysis


df['Content Rating'].value_counts()
```

Out[210]:
```
Everyone           7303
Teen               1055
Mature 17+          458
Everyone 10+        388
Adults only 18+       3
Unrated               1
Name: Content Rating, dtype: int64
```

In [ ]:
```
#Adults only 18+ and Unrated have very few records 4 only
#So we we can drop them
```

In [218]:
```
df[df['Content Rating']    'Adults only 18+'].index
```

jupyter Python_Assignment_1 Last Checkpoint: 01/29/2021 (autosaved) Logout

File Edit View Insert Cell Kernel Widgets Help Trusted Python 3

Run Markdown

In [218]:
```
df[df['Content Rating'] == 'Adults only 18+'].index
```

Out[218]:
```
Int64Index([298, 3043, 6424], dtype='int64')
```

In [219]:
```
df.drop(df[df['Content Rating'] == 'Adults only 18+'].index,axis=0,inplace=True)
```

In [220]:
```
df['Content Rating'].value_counts()
```

Out[220]:
```
Everyone        7303
Teen            1055
Mature 17+       458
Everyone 10+     388
Unrated            1
Name: Content Rating, dtype: int64
```

In [221]:
```
df[df['Content Rating'] == 'Unrated'].index
```

Out[221]:
```
Int64Index([8266], dtype='int64')
```

In [222]:
```
df.drop(df[df['Content Rating'] == 'Unrated'].index,axis=0,inplace=True)
```

In [223]:
```
df.shape
```

Out[223]:
```
(9204, 13)
```

In [227]:
```
##7. Effect of size on rating
     #a. Make a joinplot to understand the effect of size on rating
     #b. Do you see any patterns?
     #c. How do you explain the pattern?
```

Inbox (1,415) - prabhasbablu123 | Downloads/ | Python_Assignment_1 - Jupyter | Learning Hub | Learner | Board In

localhost:8888/notebooks/Downloads/Python_Assignment_1.ipynb

jupyter Python_Assignment_1 Last Checkpoint: 01/29/2021 (autosaved)

Logout

File Edit View Insert Cell Kernel Widgets Help

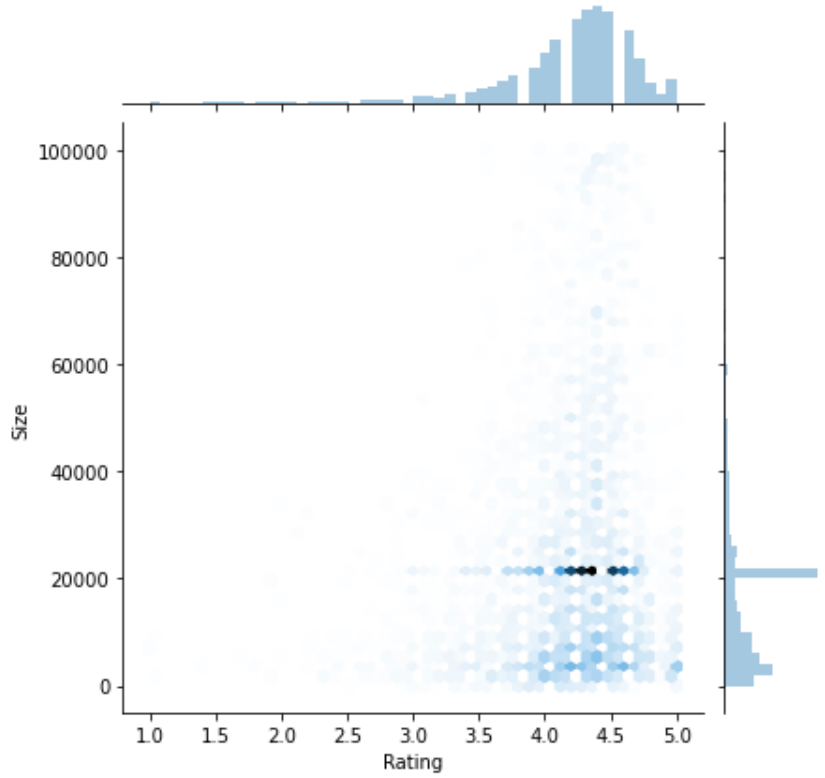
Trusted | Python 3

Run | Markdown

```
    #b. Do you see any patterns?
    #c. How do you explain the pattern?

sns.jointplot(y='Size',x='Rating',data=df, kind='hex')
```

Out[227]: <seaborn.axisgrid.JointGrid at 0x23676a1bc10>



Type here to search

ENG INTL 9:20 PM 15-Feb-21

Inbox (1,415) - prabhasbablu123 | Downloads/ | Python_Assignment_1 - Jupyter | Learning Hub | Learner | Board In

localhost:8888/notebooks/Downloads/Python_Assignment_1.ipynb

jupyter Python_Assignment_1 Last Checkpoint: 01/29/2021 (autosaved)

Logout

File Edit View Insert Cell Kernel Widgets Help

Trusted | Python 3

Run | Markdown

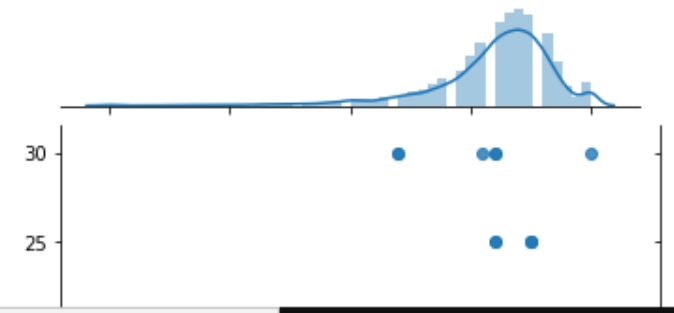
In [ ]:
```
#i dont think there is a relation between Size and Rating
# most of the concentration is in betwwen 4-4.5 in Rating
#and most of the app size is nearly 20MB
```

In [ ]:
```
##8. Effect of price on rating
  #a. Make a jointplot (with regression line)
  #b. What pattern do you see?
  #c. How do you explain the pattern?
  #d. Replot the data, this time with only records with price > 0
  #e. Does the pattern change?
  #f. What is your overall inference on the effect of price on the rating
```

In [232]:
```
sns.jointplot(x='Rating',y='Price',data=df, kind='reg')
```

```
C:\Users\vikas\anaconda3\lib\site-packages\seaborn\distributions.py:369: UserWarning: Default bandwidth for data is 0; skipping density estimation.
  warnings.warn(msg, UserWarning)
```

Out[232]: <seaborn.axisgrid.JointGrid at 0x23679095f70>



Type here to search

ENG INTL 9:20 PM 15-Feb-21

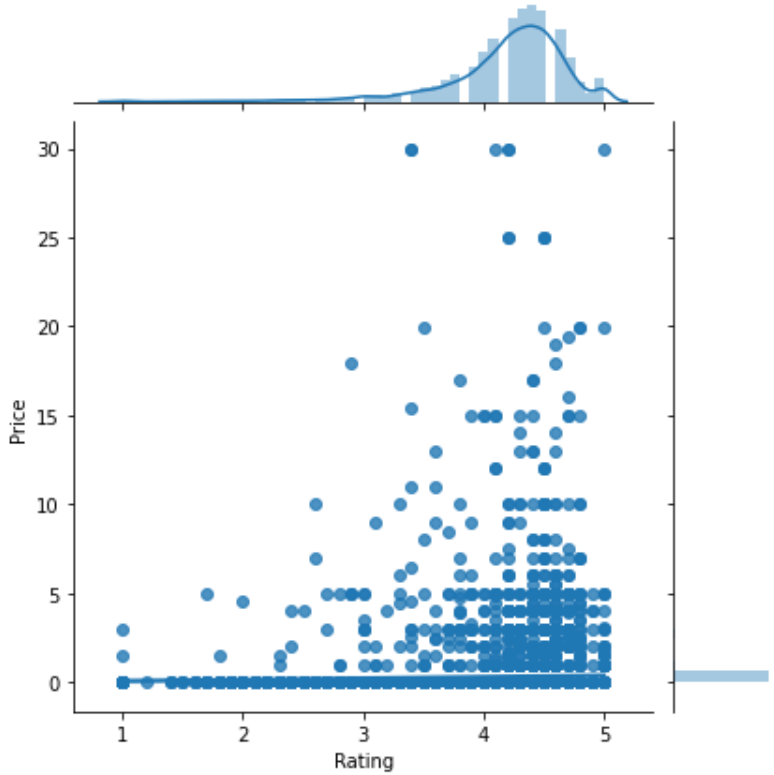
jupyter Python_Assignment_1 Last Checkpoint: 01/29/2021 (autosaved) Logout

File Edit View Insert Cell Kernel Widgets Help Trusted Python 3

Run Markdown

```
C:\Users\vikas\anaconda3\lib\site-packages\seaborn\distributions.py:369: UserWarning: Default bandwidth for data is 0; skipping density estimation.
  warnings.warn(msg, UserWarning)
```

Out[232]: <seaborn.axisgrid.JointGrid at 0x23679095f70>



jupyter Python_Assignment_1 Last Checkpoint: 01/29/2021 (autosaved) Logout

File Edit View Insert Cell Kernel Widgets Help Trusted Python 3

Run Markdown

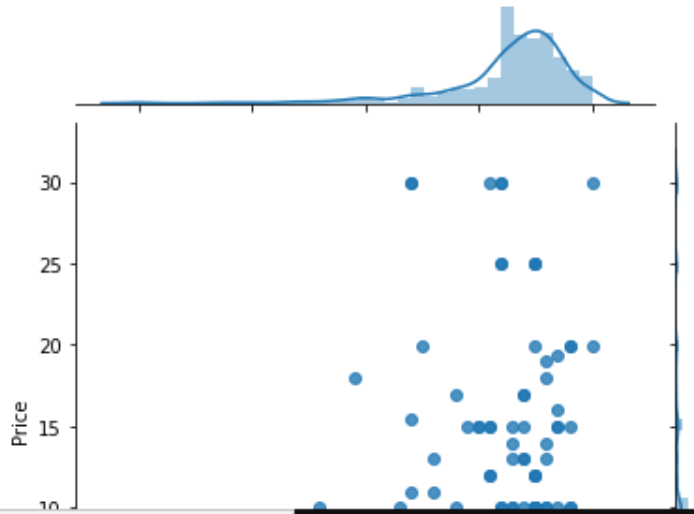
1 2 3 4 5
Rating

In [ ]:
```
#lot of apps have 0 price that means free apps
#lots of data spread around 0-5 in price
#In this range we can find 1-5 rating
#the reg line is in completely horizontal passing trough 0, so we cant say there a relation between this two columns
```

In [233]:
```
df2 = df.loc[df.Price>0]
```

In [235]:
```
sns.jointplot(y='Price', x='Rating', data=df2, kind='reg')
```

Out[235]: <seaborn.axisgrid.JointGrid at 0x2367a3328b0>



Inbox (1,415) - prabhasbablu123 | Downloads/ | Python_Assignment_1 - Jupyter | Learning Hub | Learner | Board In

localhost:8888/notebooks/Downloads/Python_Assignment_1.ipynb

jupyter Python_Assignment_1 Last Checkpoint: 01/29/2021 (autosaved)

Logout

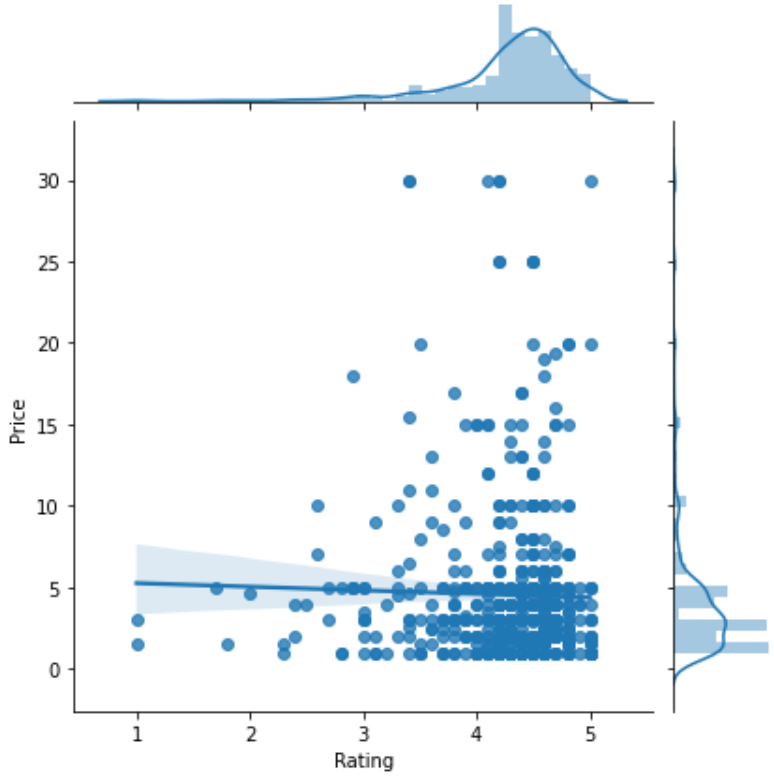
File Edit View Insert Cell Kernel Widgets Help

Trusted | Python 3

Run | Markdown

In [235]:
```
sns.jointplot(y='Price', x='Rating', data=df2, kind='reg')
```

Out[235]: <seaborn.axisgrid.JointGrid at 0x2367a3328b0>



In [ ]:
```
#Generally increasing the Prices, doesn't have signifcant effect on Higher Rating.
#For Higher Price, Rating is High and almost constant ie greater than 4
```

Type here to search

ENG INTL 9:21 PM 15-Feb-21

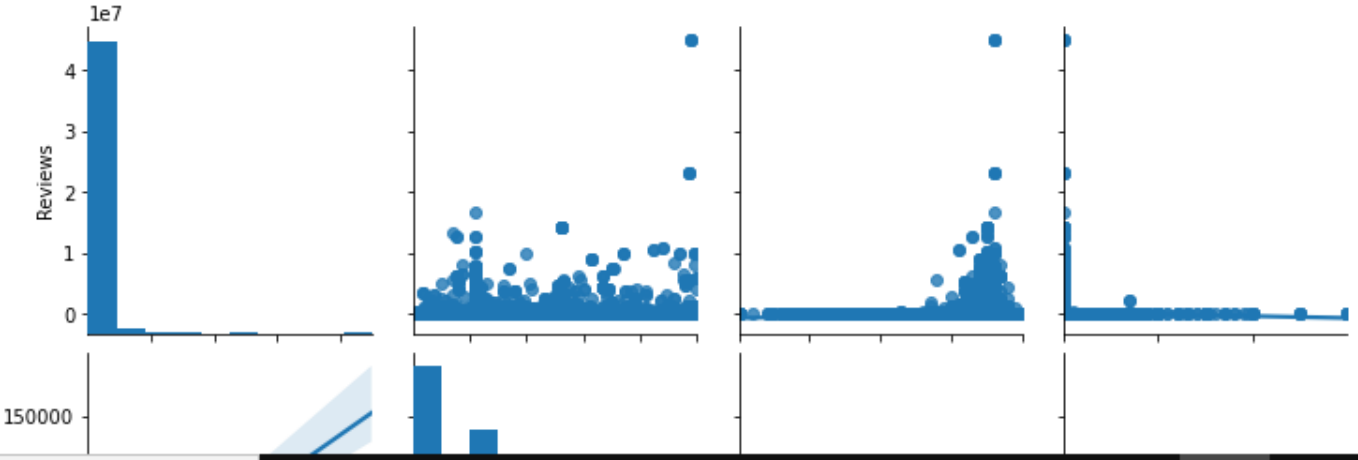
Rating

In [ ]:
```
#Generally increasing the Prices, doesn't have signifcant effect on Higher Rating.
#For Higher Price, Rating is High and almost constant ie greater than 4
#we can see there is a slightly negative relation
#which apps have 0-5 in price they have great rating like 4-5
#There is some apps which have higher in price and also in rating
#But if we ignore the data that b/w 0-5 in price we can its positively related
```

In [ ]:
```
##9. Look at all the numeric interactions together -
    #a. Make a pairplort with the colulmns - 'Reviews', 'Size', 'Rating', 'Price'
```

In [238]:
```
#sns.pairplot(df)
sns.pairplot(df,vars=['Reviews', 'Size', 'Rating', 'Price'],kind='reg')
```

Out[238]: <seaborn.axisgrid.PairGrid at 0x2367db86190>



Inbox (1,413) - prabhasbablu123 | Downloads/ | Python_Assignment_1 - Jupyter

localhost:8888/notebooks/Downloads/Python_Assignment_1.ipynb

jupyter Python_Assignment_1 Last Checkpoint: 01/29/2021 (autosaved)

Logout

File Edit View Insert Cell Kernel Widgets Help

Trusted | Python 3

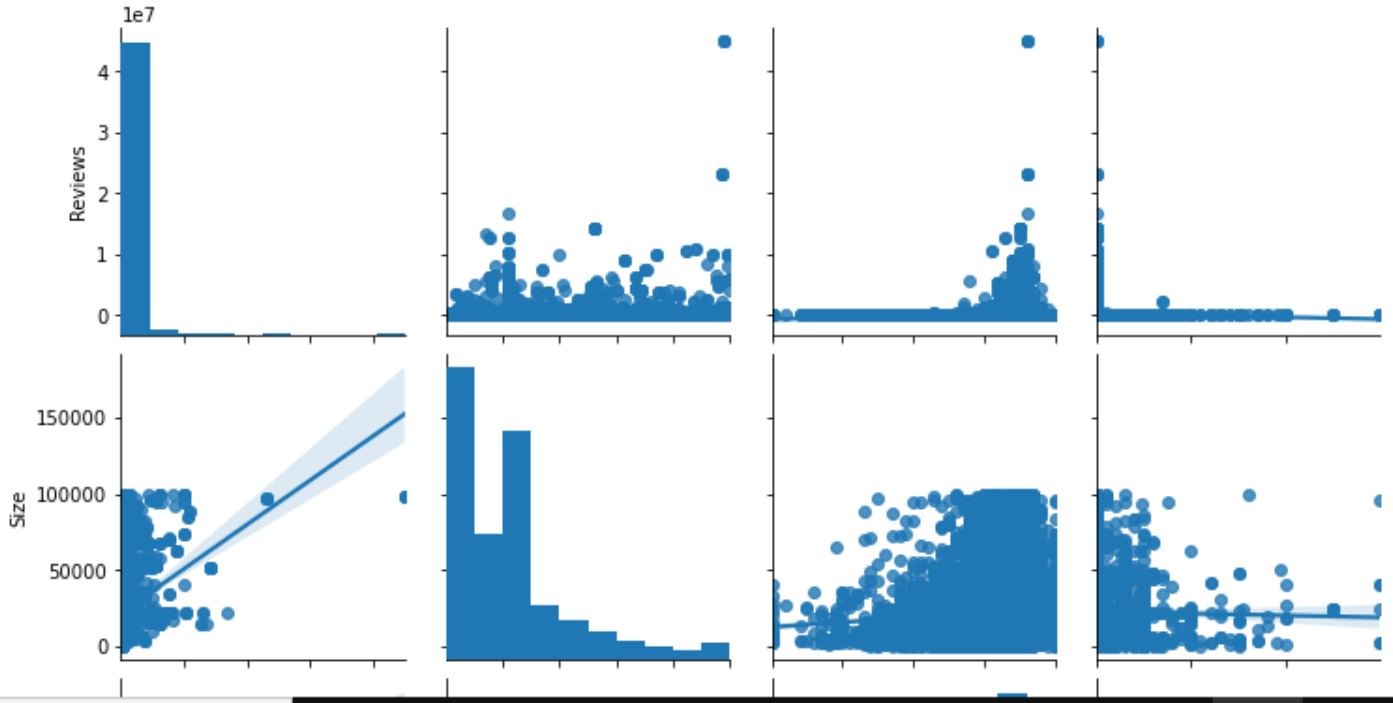
Run | Markdown

```
#But if we ignore the data that b/w 0-5 in price we can its positively related
```

In [ ]:
```
##9. Look at all the numeric interactions together -
    #a. Make a pairplort with the colulmns - 'Reviews', 'Size', 'Rating', 'Price'
```

In [238]:
```
#sns.pairplot(df)
sns.pairplot(df,vars=['Reviews', 'Size', 'Rating', 'Price'],kind='reg')
```

Out[238]: <seaborn.axisgrid.PairGrid at 0x2367db86190>



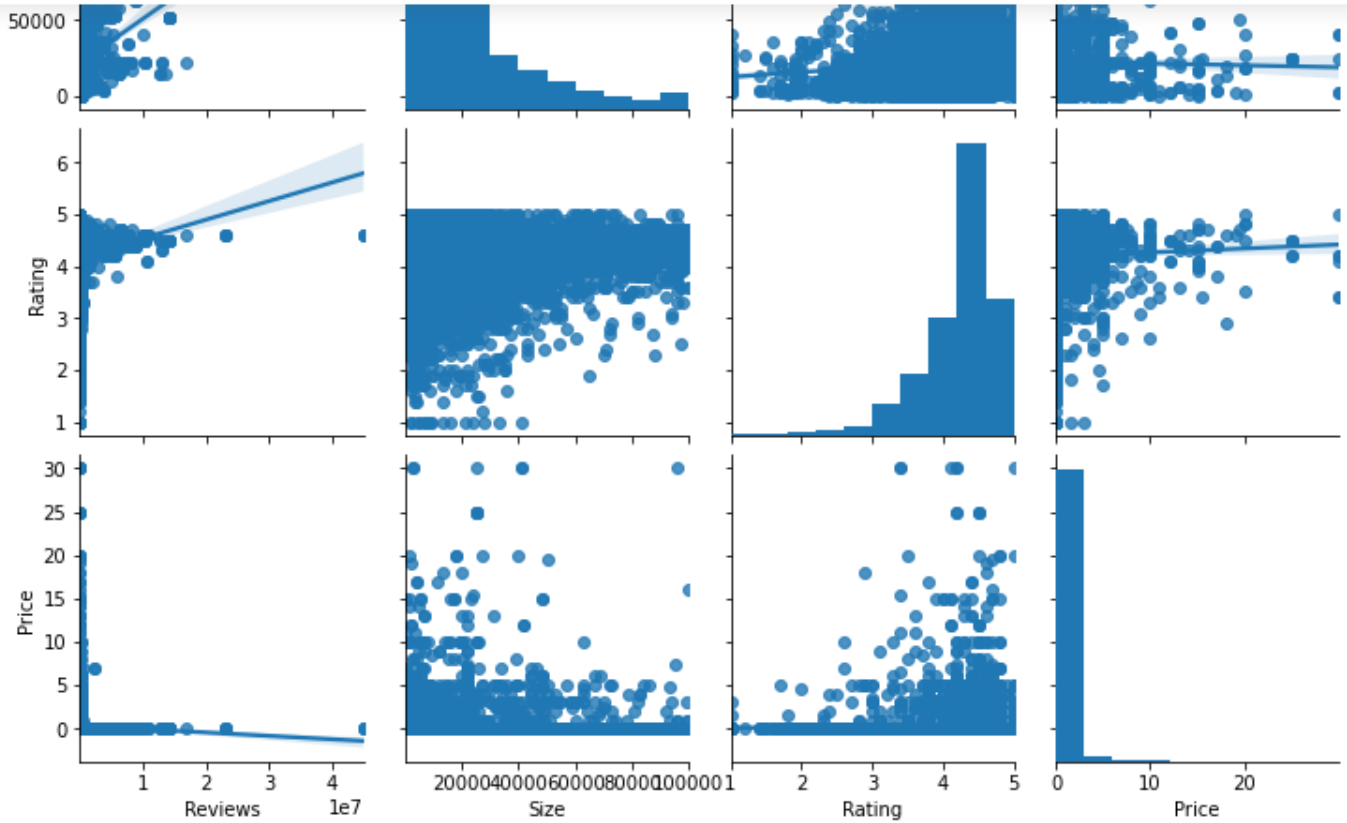
Type here to search

ENG INTL 8:52 PM 15-Feb-21

jupyter Python_Assignment_1 Last Checkpoint: 01/29/2021 (autosaved) Logout

File Edit View Insert Cell Kernel Widgets Help Trusted Python 3

Run Markdown



In [ ]:
```
#we can see that the Size and Reviews has best fit line with increasing direction
```

In [240]:
```
##10. Rating vs. content rating
   #a. Make a bar plot displaying the rating for each content rating
```

Inbox (1,413) - prabhasbablu123 | Downloads/ | Python_Assignment_1 - Jupyter

localhost:8888/notebooks/Downloads/Python_Assignment_1.ipynb

jupyter Python_Assignment_1 Last Checkpoint: 01/29/2021 (autosaved)

Logout

File Edit View Insert Cell Kernel Widgets Help

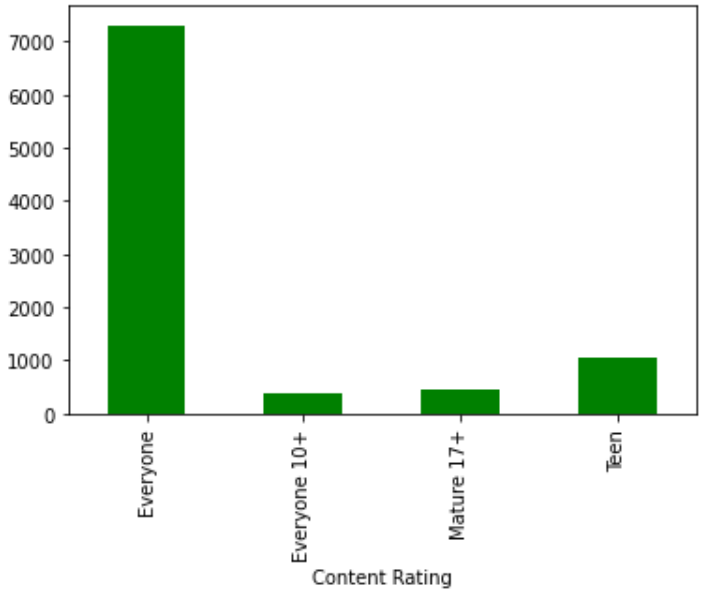
Trusted | Python 3

Run | Markdown

In [240]:
```
##10. Rating vs. content rating
   #a. Make a bar plot displaying the rating for each content rating
   #b. Which metric would you use? Mean? Median? Some other quantile?
   #c. Choose the right metric and plot
```

In [245]:
```
df.groupby(['Content Rating'])['Rating'].count().plot.bar(color="green")
```

Out[245]:
```
<matplotlib.axes._subplots.AxesSubplot at 0x2367f863160>
```



In [246]:
```
df.boxplot('Rating')
```

Type here to search

ENG INTL 8:52 PM 15-Feb-21

Inbox (1,413) - prabhasbablu123 | Downloads/ | Python_Assignment_1 - Jupyter

localhost:8888/notebooks/Downloads/Python_Assignment_1.ipynb

jupyter Python_Assignment_1 Last Checkpoint: 01/29/2021 (autosaved)

Logout

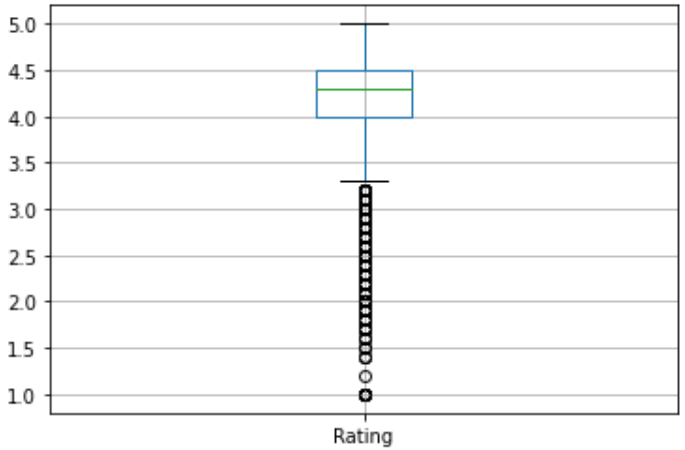
File Edit View Insert Cell Kernel Widgets Help

Trusted | Python 3

Run | Markdown

In [246]:
```
df.boxplot('Rating')
```

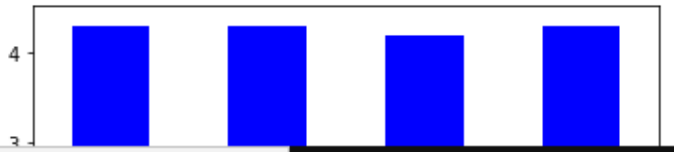
Out[246]: <matplotlib.axes._subplots.AxesSubplot at 0x2367f8c0d90>



In [ ]:
```
#there is too much outliers below min value
#we must use Median to get rid of the outliers
```

In [248]:
```
df.groupby(['Content Rating'])['Rating'].median().plot.bar(color="blue")
```

Out[248]: <matplotlib.axes._subplots.AxesSubplot at 0x2367f8c8490>



Type here to search

ENG INTL 8:52 PM 15-Feb-21

Inbox (1,413) - prabhasbablu123 | Downloads/ | Python_Assignment_1 - Jupyter

localhost:8888/notebooks/Downloads/Python_Assignment_1.ipynb

jupyter Python_Assignment_1 Last Checkpoint: 01/29/2021 (autosaved)

Logout

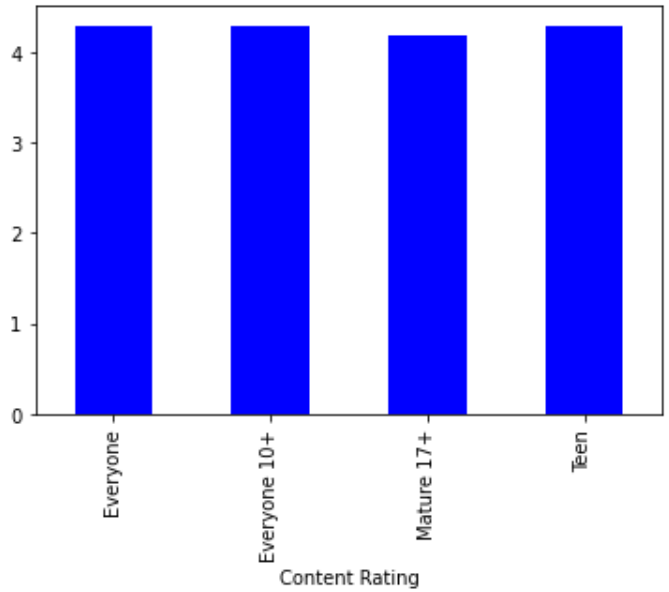
File Edit View Insert Cell Kernel Widgets Help

Trusted | Python 3

Run | Markdown

In [248]:
```
df.groupby(['Content Rating'])['Rating'].median().plot.bar(color="blue")
```

Out[248]:
```
<matplotlib.axes._subplots.AxesSubplot at 0x2367f8c8490>
```



In [ ]:
```
##11. Content rating vs. size vs. rating – 3 variables at a time
    #a. Create 5 buckets (20% records in each) based on Size
    #b. By Content Rating vs. Size buckets, get the rating (20th percentile) for each combination
    #c. Make a heatmap of this i. Annotated ii. Greens color map
    #d. What's your inference? Are lighter apps preferred in all categories? Heavier? Some?
```

Type here to search

ENG INTL 8:52 PM 15-Feb-21

Inbox (1,413) - prabhasbablu123 | Downloads/ | Python_Assignment_1 - Jupyter

localhost:8888/notebooks/Downloads/Python_Assignment_1.ipynb

jupyter Python_Assignment_1 Last Checkpoint: 01/29/2021 (autosaved)

Logout

File Edit View Insert Cell Kernel Widgets Help

Trusted | Python 3

Run | Markdown

In [ ]:
```
##11. Content rating vs. size vs. rating - 3 variables at a time
    #a. Create 5 buckets (20% records in each) based on Size
    #b. By Content Rating vs. Size buckets, get the rating (20th percentile) for each combination
    #c. Make a heatmap of this i. Annotated ii. Greens color map
    #d. What's your inference? Are lighter apps preferred in all categories? Heavier? Some?
```

In [250]:
```
bins=[0,20000,40000,60000,80000,100000]
df['size_buckets'] = pd.cut(df['Size'],bins,labels=['0-20mb','20mb-40mb','40mb-60mb','60mb-80mb','80mb-100mb'])
```

In [265]:
```
pivot=pd.pivot_table(df, values='Rating', index='size_buckets', columns='Content Rating', aggfunc=lambda x:np.quantile(x,0.2))
```

In [266]:
```
pivot
```

Out[266]:

Content Rating	Everyone	Everyone 10+	Mature 17+	Teen
size_buckets				
0-20mb	3.8	4.02	3.5	3.9
20mb-40mb	4.0	4.00	4.0	4.0
40mb-60mb	3.9	4.10	4.1	4.0
60mb-80mb	4.0	4.20	4.0	4.0
80mb-100mb	4.1	4.30	4.0	4.1

In [267]:
```
sns.heatmap(pivot, annot=True, linewidths=.5)
```

Out[267]:
```
<matplotlib.axes._subplots.AxesSubplot at 0x236009cc6d0>
```

Type here to search

ENG INTL 8:52 PM 15-Feb-21

jupyter Python_Assignment_1 Last Checkpoint: 01/29/2021 (autosaved) Logout

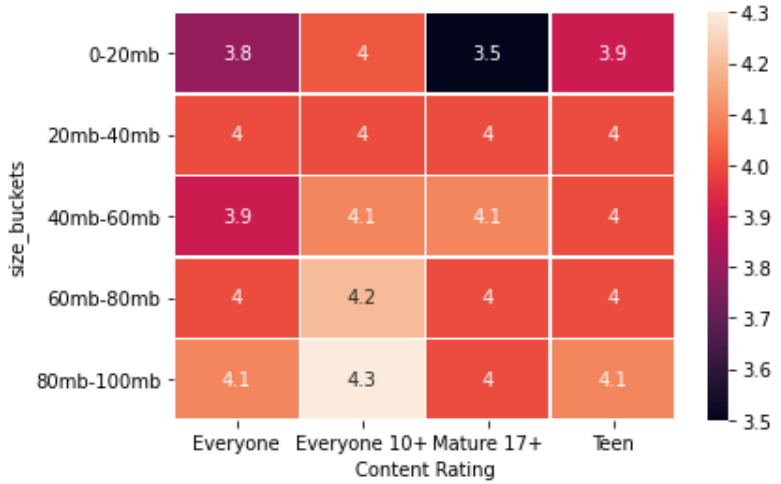
File Edit View Insert Cell Kernel Widgets Help Trusted Python 3

Run Markdown

60mb-80mb	4.0	4.20	4.0	4.0
80mb-100mb	4.1	4.30	4.0	4.1

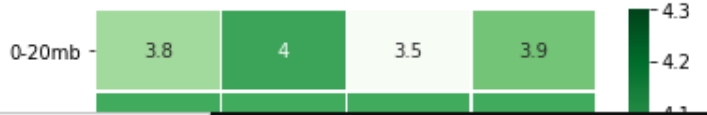
In [267]:
```
sns.heatmap(pivot, annot=True, linewidths=.5)
```

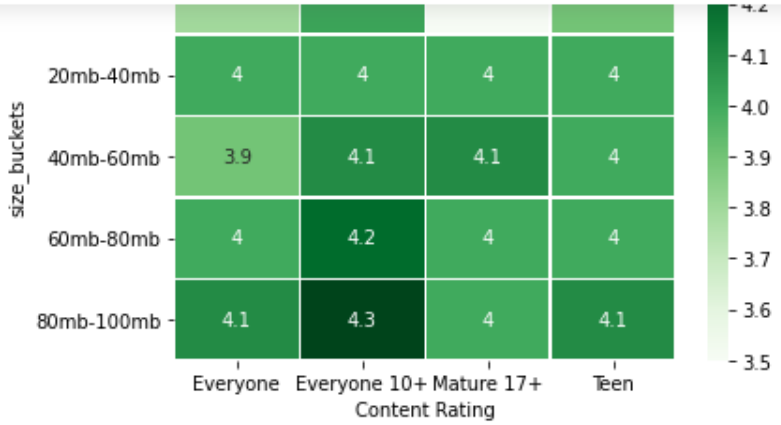
Out[267]: <matplotlib.axes._subplots.AxesSubplot at 0x236009cc6d0>



In [271]:
```
sns.heatmap(pivot, annot=True, linewidths=.5,cmap='Greens')
```

Out[271]: <matplotlib.axes._subplots.AxesSubplot at 0x236009985b0>





In [ ]:
```
#I think lighter apps prefer in all categories
#Bcoz the distribution of data lies in free apps and 0 size with rating4-5
#But nowadays games are very popular so that we can say larger apps also getting top ratings
#we can say that gaming apps like free fire ,pubg with larger size,they are getting top ratings
#but over all lighter apps prefered most us
```

THANK YOU BOARD INFINITY AND THANKS TO PUNITH SIR