In [31]:
```
import pandas as pd
import numpy as np
import matplotlib.pyplot as ply
%matplotlib inline
```

In [2]:
```
df=pd.read_csv('C:/Users/rahul/Documents/googleplaystore.csv')
df.head()
```

Out[2]:

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

In [3]:
```
df.info()
```

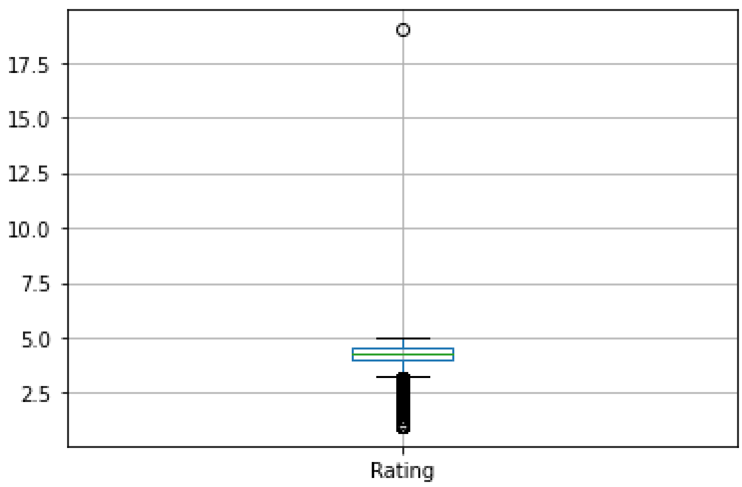
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 10841 entries, 0 to 10840
Data columns (total 13 columns):
 #   Column          Non-Null Count  Dtype  
---  ------          --------------  -----  
 0   App             10841 non-null  object 
 1   Category        10841 non-null  object 
 2   Rating          9367 non-null   float64
 3   Reviews         10841 non-null  object 
 4   Size            10841 non-null  object 
 5   Installs        10841 non-null  object 
 6   Type            10840 non-null  object 
 7   Price           10841 non-null  object 
 8   Content Rating  10840 non-null  object 
 9   Genres          10841 non-null  object 
 10  Last Updated    10841 non-null  object 
 11  Current Ver     10833 non-null  object 
 12  Android Ver     10838 non-null  object 
dtypes: float64(1), object(12)
memory usage: 1.1+ MB
```

In [6]:
```
df.isnull().sum()
```

Out[6]:
```
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 [7]:
```
#There are more than 10%
df.boxplot()
```

Out[7]:
```
<AxesSubplot:>
```



In [10]:
```
#1 outlier was observed in the ratings column and therefore the data requires cleani
df[df.Rating>5]
```

Out[10]:

	App	Category	Rating	Reviews	Size	Installs	Type	Price	Content Rating	Genres
10472	Life Made WI-Fi Touchscreen Photo Frame	1.9	19.0	3.0M	1,000+	Free	0	Everyone	NaN	February 11, 2018

In [12]:
```
df.drop([10472],inplace=True)
```

In [14]:
```
df[10471:10474]
```

Out[14]:

	App	Category	Rating	Reviews	Size	Installs	Type	Price	Content Rating	
10471	Xposed Wi-Fi-Pwd	PERSONALIZATION	3.5	1042	404k	100,000+	Free	0	Everyone	Pers

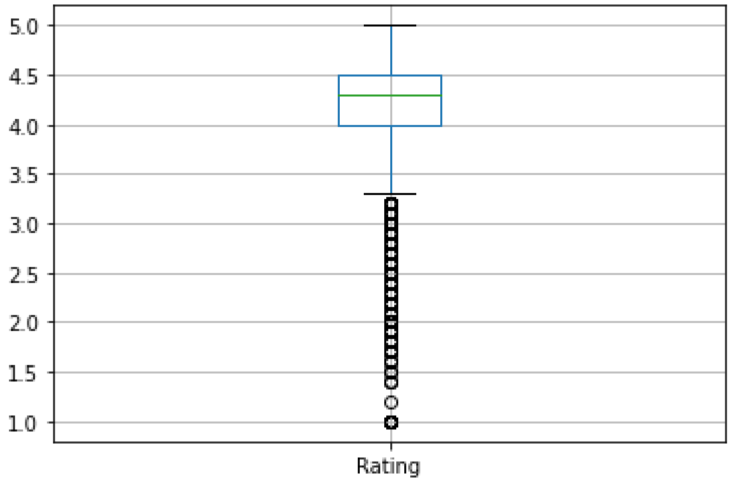
	App	Category	Rating	Reviews	Size	Installs	Type	Price	Content Rating	
10473	osmino Wi-Fi: free WiFi	TOOLS	4.2	134203	4.1M	10,000,000+	Free	0	Everyone	
10474	Sat-Fi Voice	COMMUNICATION	3.4	37	14M	1,000+	Free	0	Everyone	Com

In [15]:

```
df.boxplot()
```

Out[15]:

```
<AxesSubplot:>
```

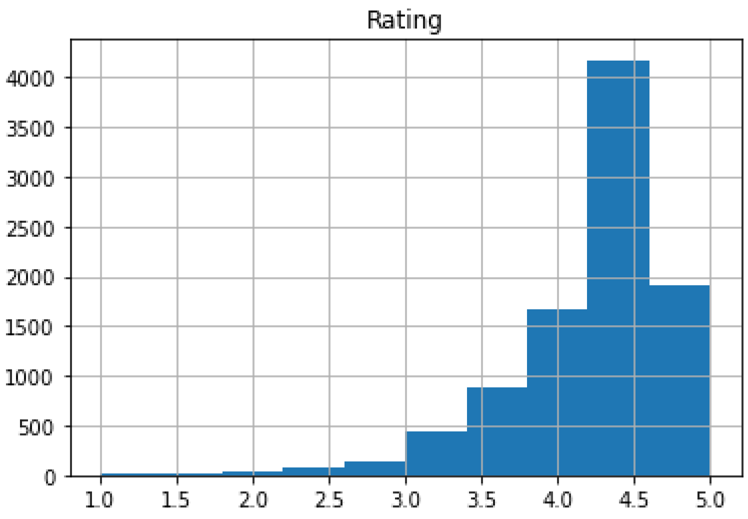


In [16]:

```
#Most of the ratings are between 4 and 4.5
df.hist()
```

Out[16]:

```
array([[<AxesSubplot:title={'center':'Rating'}>]], dtype=object)
```



In [ ]:

```
#This is a negatively skewed/ left skewed graph for the data and thus median is to b
#I shall be filling all the blank rows with median value
```

In [18]:

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

```
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 [ ]:

```
#Therefore the 1474 null values under the column rating is to be filled with the med
#The other null values under Type, Current version and android version are to be fil
```

In [19]:

```
def impute_median(series):
    return series.fillna(series.median())
```

In [30]:

```
df.Rating = df['Rating'].transform(impute_median)
df.isnull().sum()
```

Out[30]:

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

In [22]:

```
#Checking the mode for individual categories just to make sure there are no bimodal
print(df['Type'].mode())
print(df['Current Ver'].mode())
print(df['Android Ver'].mode())
```

```
0    Free
dtype: object
0    Varies with device
dtype: object
0    4.1 and up
dtype: object
```

In [21]:

```
df['Type'].fillna(str(df['Type'].mode().values[0]), inplace=True)
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)
df.isnull().sum()
```

Out[21]:

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

In [26]:

```
df['Price'] = df['Price'].apply(lambda x: str(x).replace('$', '') if '$' in str(x) e
df['Price'] = df['Price'].apply(lambda x: float(x))
df['Reviews'] = pd.to_numeric(google_data['Reviews'], errors='coerce')
#Dollar sign from the prices were removed and converted to a string.This string was
#All the reviews were subjected to to_numeric functions. All errors are to be ignore
```

In [27]:

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

Out[27]:

	App	Category	Rating	Reviews	Size	Installs	Type	Price	Content Rating	
0	Photo Editor & Candy Camera & Grid & ScrapBook	ART_AND_DESIGN	4.1	159	19M	10000.0	Free	0	Everyone	Art &
1	Coloring book moana	ART_AND_DESIGN	3.9	967	14M	500000.0	Free	0	Everyone	Design;
2	U Launcher Lite – FREE Live Cool Themes, Hide ...	ART_AND_DESIGN	4.7	87510	8.7M	5000000.0	Free	0	Everyone	Art &
3	Sketch - Draw & Paint	ART_AND_DESIGN	4.5	215644	25M	50000000.0	Free	0	Teen	Art &
4	Pixel Draw - Number Art Coloring Book	ART_AND_DESIGN	4.3	967	2.8M	100000.0	Free	0	Everyone	Design;C

In [34]:

```
df.to_csv('App_ratings_new.csv')
```

In [ ]:

```
#This new CSV file is being used for visualisation in Tableau
```