Importing Libraries and Package

In [1]:
```
import pandas as pd
import numpy as np
import seaborn as sns
import matplotlib.pyplot as plt
%matplotlib inline
```

Reading the data set

In [2]:
```
df=pd.read_csv("C://Users//N-A-N-I//Desktop//CIP//PR-1//archive//googleplays
df.head() #head()-- display the top 5 records of the data set
```

Out[2]:

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

In [3]:
```
df.tail() #tail()--- display the bottom 5 record of the data set
```

Out[3]:

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

In [4]:
```
df.shape #shape---strcture of the data set
```

Out[4]:
```
(10841, 13)
```

In [5]:
```
print("Number of Rows:",df.shape[0])
print("Number of Columns:",df.shape[1])
```

```
Number of Rows: 10841
Number of Columns: 13
```

In [6]:
```
df.dtypes #data type of each attribute
```

Out[6]:
```
App               object
Category          object
Rating           float64
Reviews           object
Size              object
Installs          object
Type              object
Price             object
Content Rating    object
Genres            object
Last Updated      object
Current Ver       object
Android Ver       object
dtype: object
```

In [7]:
```
df.info() #full infromation of the data set
```

```
<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 [8]:
```
df.describe() #It is displayed stastics methods of the data set
```

Out[8]:

	Rating
count	9367.000000
mean	4.193338
std	0.537431
min	1.000000
25%	4.000000
50%	4.300000
75%	4.500000
max	19.000000

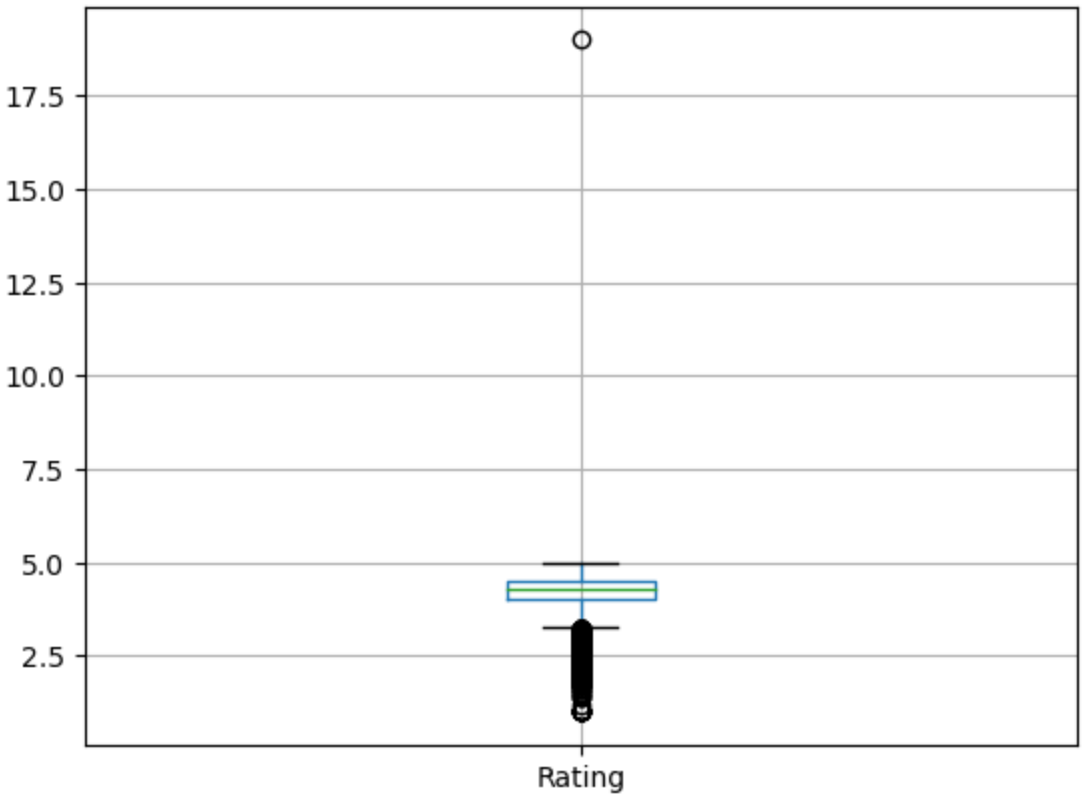
In [9]:
```
df.describe(include="all")
```

Out[9]:

	App	Category	Rating	Reviews	Size	Installs	Type	Price	Co R
count	10841	10841	9367.000000	10841	10841	10841	10840	10841	
unique	9660	34	NaN	6002	462	22	3	93	
top	ROBLOX	FAMILY	NaN	0	Varies with device	1,000,000+	Free	0	Eve
freq	9	1972	NaN	596	1695	1579	10039	10040	
mean	NaN	NaN	4.193338	NaN	NaN	NaN	NaN	NaN	
std	NaN	NaN	0.537431	NaN	NaN	NaN	NaN	NaN	
min	NaN	NaN	1.000000	NaN	NaN	NaN	NaN	NaN	
25%	NaN	NaN	4.000000	NaN	NaN	NaN	NaN	NaN	
50%	NaN	NaN	4.300000	NaN	NaN	NaN	NaN	NaN	
75%	NaN	NaN	4.500000	NaN	NaN	NaN	NaN	NaN	
max	NaN	NaN	19.000000	NaN	NaN	NaN	NaN	NaN	

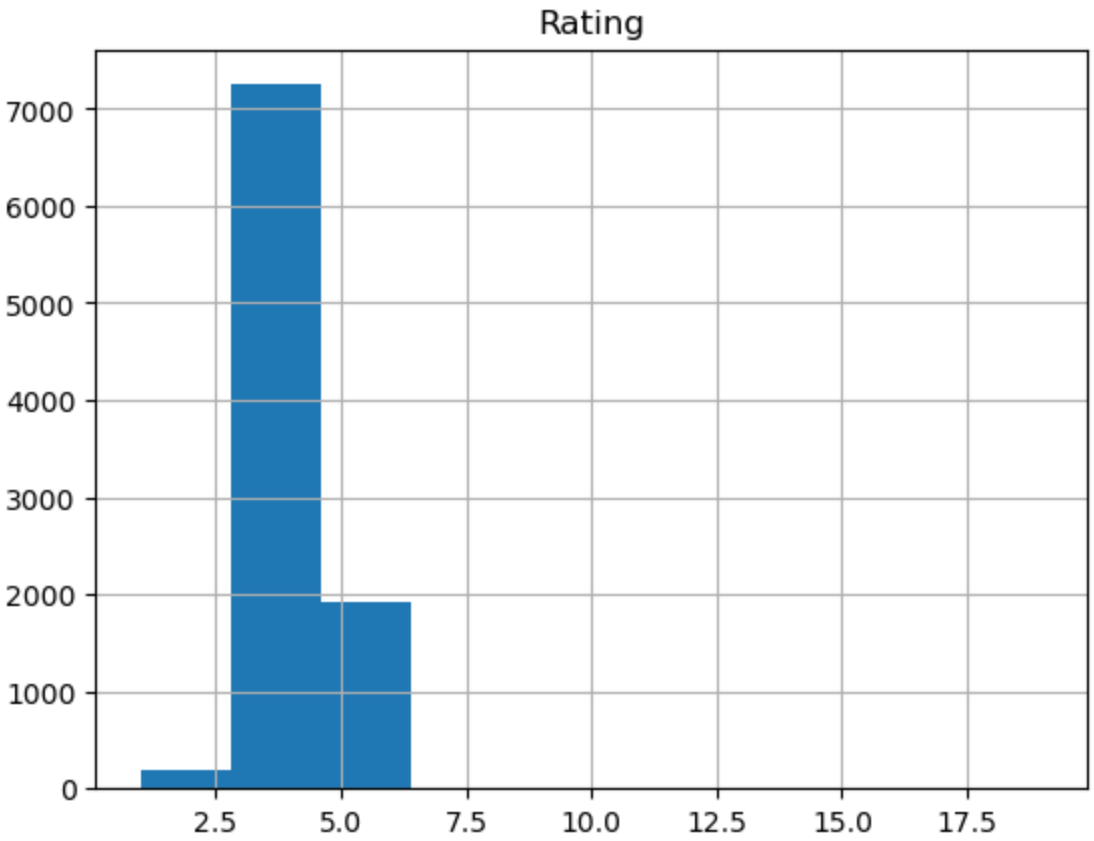
In [10]:
```
df.boxplot()  # a special type of diagram that shows the quartiles in a box
              #the line extending from the lowest to the highest value.
```

Out[10]: <AxesSubplot:>



In [11]:
```
df.hist() #A histogram is a graphical representation of data points
```

Out[11]: array([[<AxesSubplot:title={'center':'Rating'}>]], dtype=object)



Data Cleaning

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]:
```
df.isnull()
```

Out[13]:

	App	Category	Rating	Reviews	Size	Installs	Type	Price	Content Rating	Genres
0	False	False	False	False	False	False	False	False	False	False
1	False	False	False	False	False	False	False	False	False	False
2	False	False	False	False	False	False	False	False	False	False
3	False	False	False	False	False	False	False	False	False	False
4	False	False	False	False	False	False	False	False	False	False
...	...	...	...	...	...	...	...	...	...	...
10836	False	False	False	False	False	False	False	False	False	False
10837	False	False	False	False	False	False	False	False	False	False
10838	False	False	True	False	False	False	False	False	False	False
10839	False	False	False	False	False	False	False	False	False	False
10840	False	False	False	False	False	False	False	False	False	False

10841 rows × 13 columns

In [14]:
```
df[df.Rating>5]
```

Out[14]:

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

In [15]:
```
df.drop([10472],inplace=True)
```

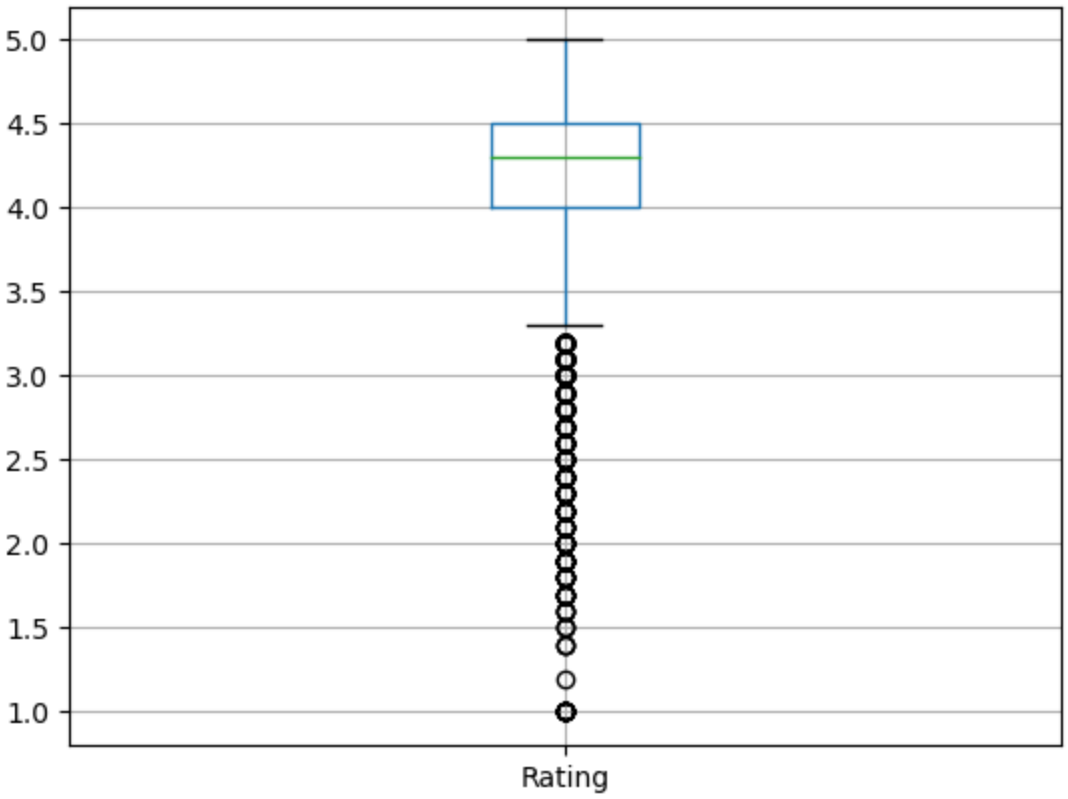
In [16]:
```
df[10470:10475]
```

Out[16]:

	App	Category	Rating	Reviews	Size	Installs	Type	Price	
10470	Jazz Wi-Fi	COMMUNICATION	3.4	49	4.0M	10,000+	Free	0	E
10471	Xposed Wi-Fi-Pwd	PERSONALIZATION	3.5	1042	404k	100,000+	Free	0	E
10473	osmino Wi-Fi: free WiFi	TOOLS	4.2	134203	4.1M	10,000,000+	Free	0	E
10474	Sat-Fi Voice	COMMUNICATION	3.4	37	14M	1,000+	Free	0	E
10475	Wi-Fi Visualizer	TOOLS	3.9	132	2.6M	50,000+	Free	0	E

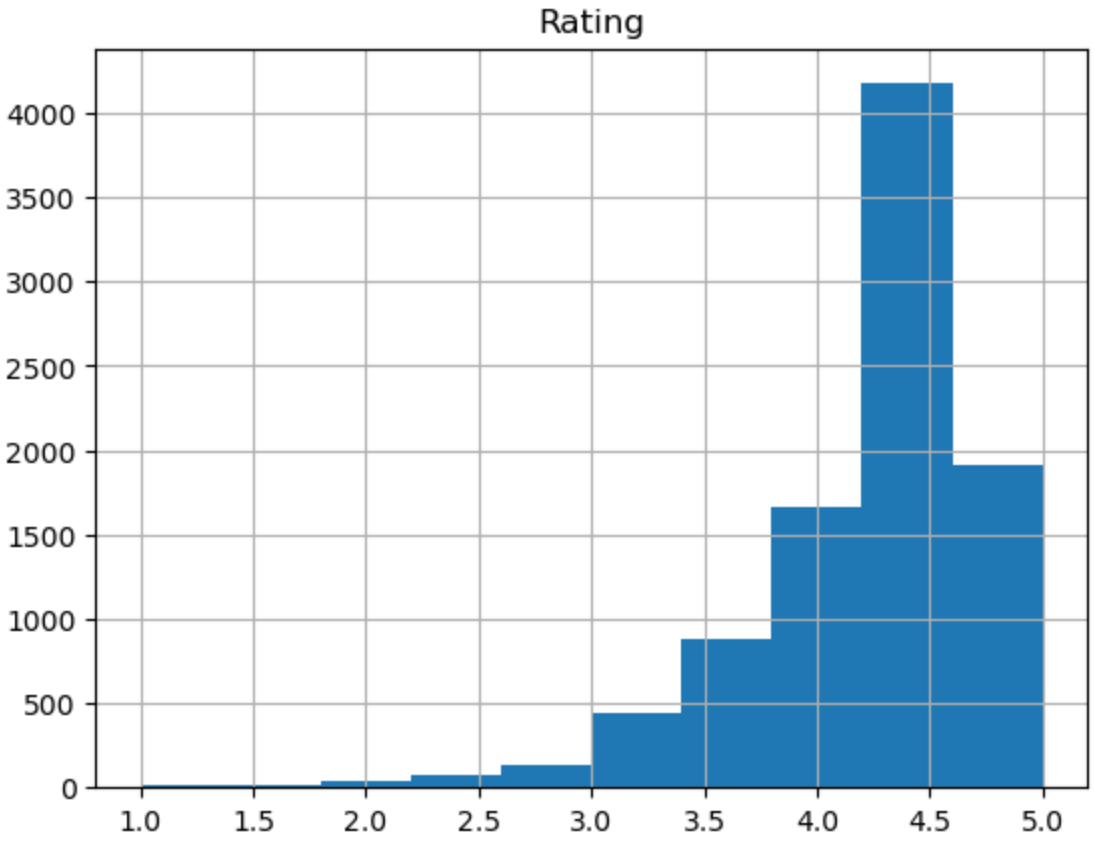
In [17]:
```
df.boxplot()
```

Out[17]: <AxesSubplot:>



In [18]:
```
df.hist()
```

Out[18]: array([[<AxesSubplot:title={'center':'Rating'}>]], dtype=object)



In [19]:
```
#define Function Here....medianvalues


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

In [20]:
```
df['Rating']=df['Rating'].transform(impute_median)
```

In [21]:
```
df.isnull().sum()
```

Out[21]:
```
App               0
Category          0
Rating            0
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 [22]:
```
#mode for categorical values


e'].mode())
```

```
print(df['Current Ver'].mode())
print(df['Android Ver'].mode())
```

```
0    Free
Name: Type, dtype: object
0    Varies with device
Name: Current Ver, dtype: object
0    4.1 and up
Name: Android Ver, dtype: object
```

In [23]:
```
df['Type'].fillna(str(df['Type'].mode().values[0]),inplace=True)
df['Current Ver'].fillna(str(df['Current Ver'].mode().values[0]),inplace=Tru
df['Android Ver'].fillna(str(df['Android Ver'].mode().values[0]),inplace=Tru
```

In [24]:
```
df.isnull().sum()
```

Out[24]:
```
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 [25]:
```
df.head()
```

Out[25]:

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

In [26]:
```
df['Reviews']=df['Reviews'].apply(lambda x:float(x))
df['Installs'] = df['Installs'].apply(lambda x: str(x).replace('+','').repla
df['Price'] = df['Price'].apply(lambda x: str(x).replace('$','')).astype('fl
```

In [27]:
```
df.head()
```

Out[27]:

	App	Category	Rating	Reviews	Size	Installs	Type	Price	Conte Ratir
0	Photo Editor & Candy Camera & Grid & ScrapBook	ART_AND_DESIGN	4.1	159.0	19M	10000	Free	0.0	Everyor
1	Coloring book moana	ART_AND_DESIGN	3.9	967.0	14M	500000	Free	0.0	Everyor
2	U Launcher Lite – FREE Live Cool Themes, Hide ...	ART_AND_DESIGN	4.7	87510.0	8.7M	5000000	Free	0.0	Everyor
3	Sketch - Draw & Paint	ART_AND_DESIGN	4.5	215644.0	25M	50000000	Free	0.0	Tee
4	Pixel Draw - Number Art Coloring Book	ART_AND_DESIGN	4.3	967.0	2.8M	100000	Free	0.0	Everyor

In [28]:
```
df.describe()
```

Out[28]:

	Rating	Reviews	Installs	Price
count	10840.000000	1.084000e+04	1.084000e+04	10840.000000
mean	4.206476	4.441529e+05	1.546434e+07	1.027368
std	0.480342	2.927761e+06	8.502936e+07	15.949703
min	1.000000	0.000000e+00	0.000000e+00	0.000000
25%	4.100000	3.800000e+01	1.000000e+03	0.000000
50%	4.300000	2.094000e+03	1.000000e+05	0.000000
75%	4.500000	5.477550e+04	5.000000e+06	0.000000
max	5.000000	7.815831e+07	1.000000e+09	400.000000

Data Visualization

In [29]:
```
rama = df.groupby('Category')
x=rama['Rating'].mean()
y=rama['Price'].sum()
z=rama['Reviews'].max()
```

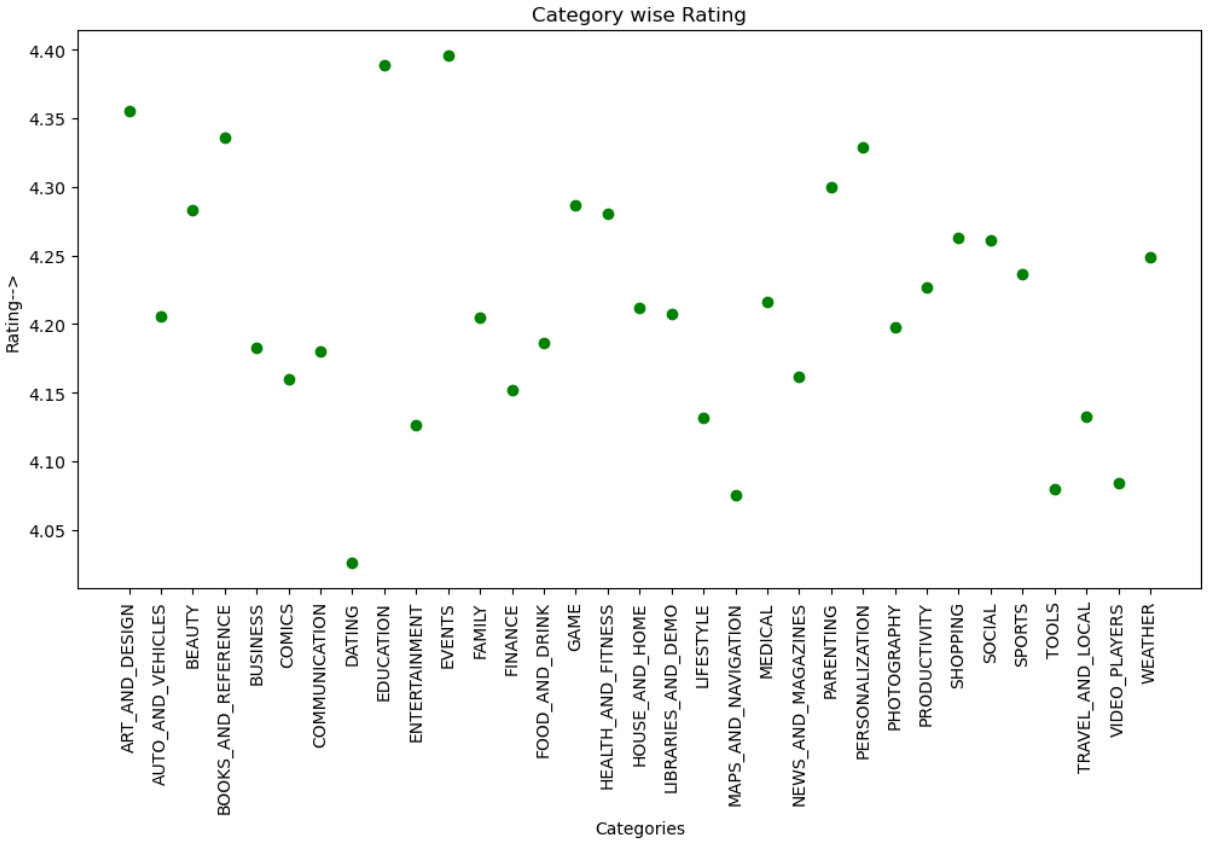
```
print(x)
print(y)
print(z)
```

```
Category
ART_AND_DESIGN         4.355385
AUTO_AND_VEHICLES      4.205882
BEAUTY                 4.283019
BOOKS_AND_REFERENCE    4.335498
BUSINESS               4.182391
COMICS                 4.160000
COMMUNICATION          4.180103
DATING                 4.025641
EDUCATION              4.388462
ENTERTAINMENT          4.126174
EVENTS                 4.395313
FAMILY                 4.204564
FINANCE                4.151639
FOOD_AND_DRINK         4.185827
GAME                   4.286888
HEALTH_AND_FITNESS     4.280059
HOUSE_AND_HOME         4.211364
LIBRARIES_AND_DEMO     4.207059
LIFESTYLE              4.131414
MAPS_AND_NAVIGATION    4.075182
MEDICAL                4.216199
NEWS_AND_MAGAZINES     4.161837
PARENTING              4.300000
PERSONALIZATION        4.328827
PHOTOGRAPHY            4.197910
PRODUCTIVITY           4.226651
SHOPPING               4.263077
SOCIAL                 4.261017
SPORTS                 4.236458
TOOLS                  4.080071
TRAVEL_AND_LOCAL       4.132946
VIDEO_PLAYERS          4.084000
WEATHER                4.248780
Name: Rating, dtype: float64
Category
ART_AND_DESIGN           5.97
AUTO_AND_VEHICLES       13.47
BEAUTY                   0.00
BOOKS_AND_REFERENCE    119.77
BUSINESS               185.27
COMICS                   0.00
COMMUNICATION           83.14
DATING                  31.43
EDUCATION               17.96
ENTERTAINMENT            7.98
EVENTS                 109.99
FAMILY                2434.78
FINANCE               2900.83
FOOD_AND_DRINK           8.48
GAME                   287.30
HEALTH_AND_FITNESS      67.34
HOUSE_AND_HOME           0.00
LIBRARIES_AND_DEMO       0.99
LIFESTYLE             2360.87
                        26.95
```

```
MEDICAL                1439.96
NEWS_AND_MAGAZINES        3.98
PARENTING                 9.58
PERSONALIZATION         153.96
PHOTOGRAPHY             134.21
PRODUCTIVITY            250.93
SHOPPING                  5.48
SOCIAL                   15.97
SPORTS                  100.00
TOOLS                   267.25
TRAVEL_AND_LOCAL         49.95
VIDEO_PLAYERS            10.46
WEATHER                  32.42
Name: Price, dtype: float64
Category
ART_AND_DESIGN           295237.0
AUTO_AND_VEHICLES        271920.0
BEAUTY                   113715.0
BOOKS_AND_REFERENCE     2915189.0
BUSINESS                1279800.0
COMICS                  1013944.0
COMMUNICATION          69119316.0
DATING                   516917.0
EDUCATION               6290507.0
ENTERTAINMENT           7165362.0
EVENTS                    40113.0
FAMILY                 44881447.0
FINANCE                 1374549.0
FOOD_AND_DRINK          1032935.0
GAME                   44893888.0
HEALTH_AND_FITNESS      4559407.0
HOUSE_AND_HOME           417907.0
LIBRARIES_AND_DEMO       332083.0
LIFESTYLE               2789775.0
MAPS_AND_NAVIGATION     7232629.0
MEDICAL                  156410.0
NEWS_AND_MAGAZINES     11667403.0
PARENTING                658087.0
PERSONALIZATION         7464996.0
PHOTOGRAPHY            10859051.0
PRODUCTIVITY            5383985.0
SHOPPING                6212081.0
SOCIAL                 78158306.0
SPORTS                 14184910.0
TOOLS                  42916526.0
TRAVEL_AND_LOCAL        9235373.0
VIDEO_PLAYERS          25655305.0
WEATHER                 2371543.0
Name: Reviews, dtype: float64
```

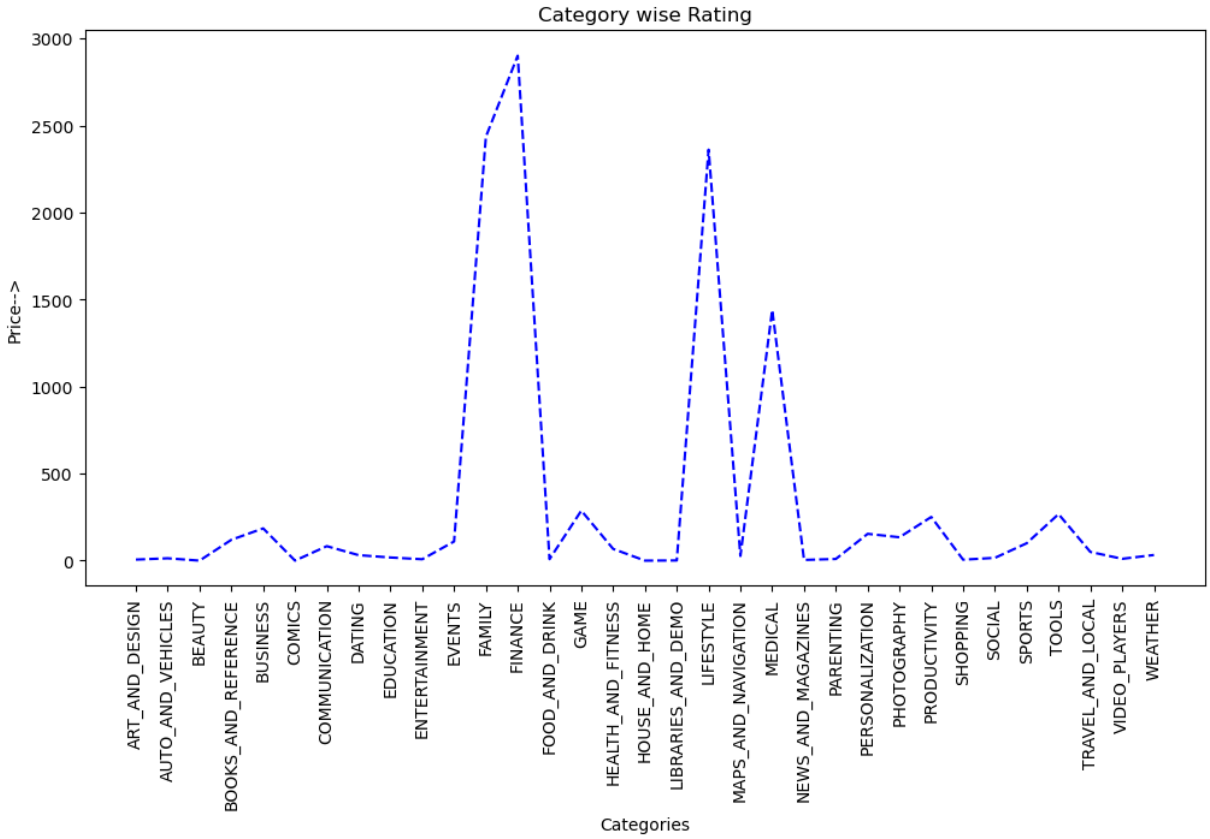
In [30]:
```
plt.figure(figsize=(12,6))
plt.plot(x,'ro',color='g');
plt.xticks(rotation=90);
plt.title('Category wise Rating');
plt.xlabel('Categories');
plt.ylabel('Rating-->');
```

```
C:\Users\N-A-N-I\AppData\Local\Temp\ipykernel_3404\3367091997.py:2: UserWarni
ng: color is redundantly defined by the 'color' keyword argument and the fmt
string "ro" (-> color='r'). The keyword argument will take precedence.
  plt.plot(x,'ro',color='g');
```



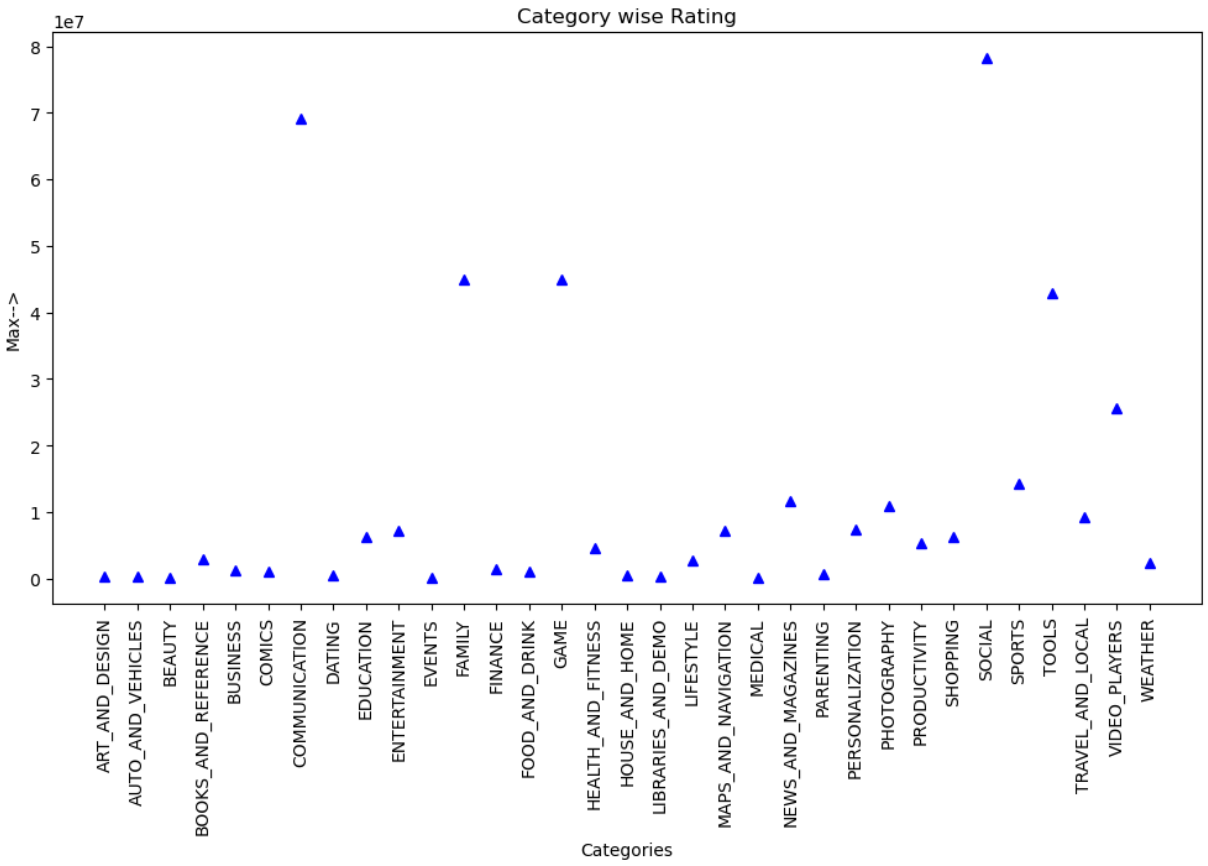
In [31]:
```
plt.figure(figsize=(12,6))
plt.plot(y,'r--',color='b');
plt.xticks(rotation=90);
plt.title('Category wise Rating');
plt.xlabel('Categories');
plt.ylabel('Price-->');
```

```
C:\Users\N-A-N-I\AppData\Local\Temp\ipykernel_3404\1018834857.py:2: UserWarni
ng: color is redundantly defined by the 'color' keyword argument and the fmt
string "r--" (-> color='r'). The keyword argument will take precedence.
  plt.plot(y,'r--',color='b');
```



In [32]:
```
plt.figure(figsize=(12,6))
plt.plot(z,'g^',color='b');
plt.xticks(rotation=90);
plt.title('Category wise Rating');
plt.xlabel('Categories');
plt.ylabel('Max-->');
```

```
C:\Users\N-A-N-I\AppData\Local\Temp\ipykernel_3404\2806569214.py:2: UserWarni
ng: color is redundantly defined by the 'color' keyword argument and the fmt
string "g^" (-> color='g'). The keyword argument will take precedence.
  plt.plot(z,'g^',color='b');
```



In [ ]:

In [ ]: