In [1]:

```
# import libraries
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
import numpy as nm
plt.figure(figsize=(80,50))

# load data
df=pd.read_csv(r"C:\Users\HARDIK\Desktop\movies.csv")
```

```
<Figure size 8000x5000 with 0 Axes>
```

In [2]:

```
df.head(3)
```

Out[2]:

	name	rating	genre	year	released	score	votes	director	writer	
0	The Shining	R	Drama	1980	June13,1980	8.4	927000.0	Stanley Kubrick	Stephen King	Nicho
1	The Blue Lagoon	R	Adventure	1980	July2,1980	5.8	65000.0	Randal Kleiser	Henry De Vere Stacpoole	Bro Shi
2	Star Wars: Episode V - The Empire Strikes Back	PG	Action	1980	June20,1980	8.7	1200000.0	Irvin Kershner	Leigh Brackett	M Ha

In [3]:

```
#data cleaning
df.isnull().sum()
```

Out[3]:

```
name           0
rating        77
genre          0
year           0
released       2
score          3
votes          3
director       0
writer         3
star           1
country        3
budget      2171
gross        189
company       17
runtime        4
dtype: int64
```

In [4]:

```
df['budget']=df['budget'].fillna(df['budget'].mean())
df['gross']=df['gross'].fillna(df['gross'].mean())
df['votes']=df['votes'].fillna(df['votes'].mean())
df['score']=df['score'].fillna(df['score'].mean())
df.dropna(inplace=True)
```

In [5]:

```
#changing data types
df.dtypes

df['budget']=df['budget'].astype('int64')
df['gross']=df['gross'].astype('int64')
df['runtime']=df['runtime'].astype('int64')
df.head(3)
```

Out[5]:

	name	rating	genre	year	released	score	votes	director	writer	
0	The Shining	R	Drama	1980	June13,1980	8.4	927000.0	Stanley Kubrick	Stephen King	J Nicho
1	The Blue Lagoon	R	Adventure	1980	July2,1980	5.8	65000.0	Randal Kleiser	Henry De Vere Stacpoole	Bro Shi
2	Star Wars: Episode V - The Empire Strikes Back	PG	Action	1980	June20,1980	8.7	1200000.0	Irvin Kershner	Leigh Brackett	M Ha

In [6]:

```
df['correct_year']=df['released'].astype('str').str[-4:]
```

In [7]:

```
pd.set_option('display.max_rows',None)
```

In [8]:

```
df=df.sort_values(by=['gross'],inplace=False,ascending=False)
df.head(3)
```

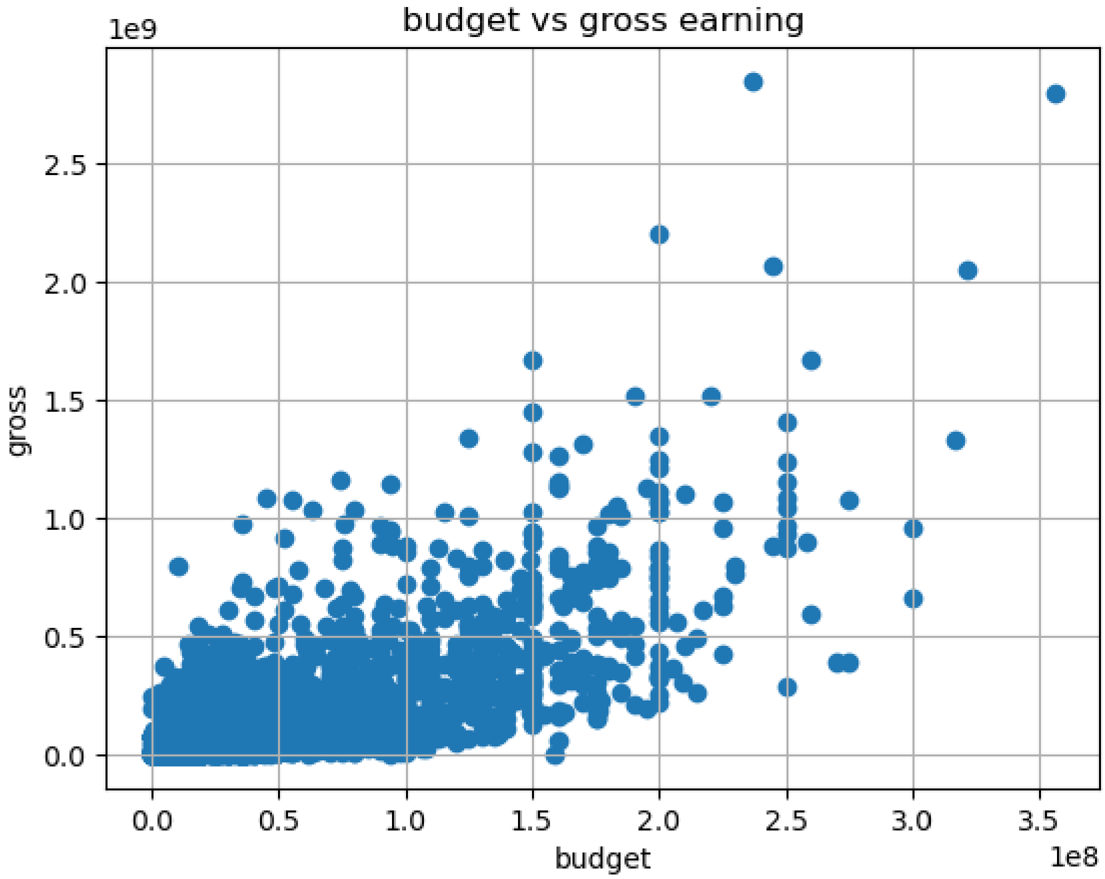
Out[8]:

	name	rating	genre	year	released	score	votes	director	write
5445	Avatar	PG-13	Action	2009	December18,2009	7.8	1100000.0	James Cameron	Jame Camero
7445	Avengers: Endgame	PG-13	Action	2019	April26,2019	8.4	903000.0	Anthony Russo	Christophe Marku
3045	Titanic	PG-13	Drama	1997	December19,1997	7.8	1100000.0	James Cameron	Jame Camero

In [9]:

```
# comparing budget and gross with scatter plot
plt.scatter(x=df['budget'],y=df['gross'])
plt.title('budget vs gross earning')

plt.xlabel('budget')
plt.ylabel('gross')
plt.grid()
plt.show()
```



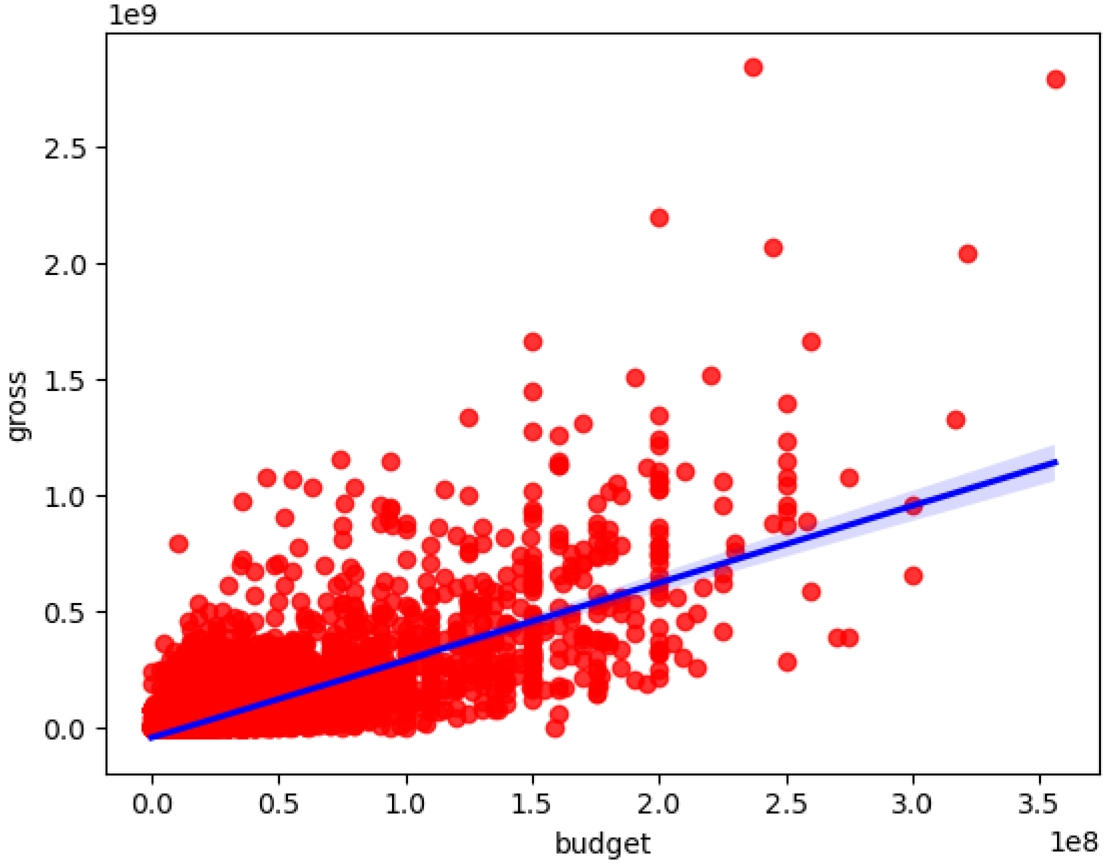
In [10]:

```
# comparing budget and gross with seaborn

sns.regplot(x='budget',y='gross',data=df,scatter_kws={"color":"red"},line_kws={"color":"b
```

Out[10]:

```
<Axes: xlabel='budget', ylabel='gross'>
```



In [11]:

```
df.corr(method='pearson')
```

```
C:\Users\HARDIK\AppData\Local\Temp\ipykernel_3660\1928163937.py:1: FutureW
arning: The default value of numeric_only in DataFrame.corr is deprecated.
In a future version, it will default to False. Select only valid columns o
r specify the value of numeric_only to silence this warning.
  df.corr(method='pearson')
```

Out[11]:

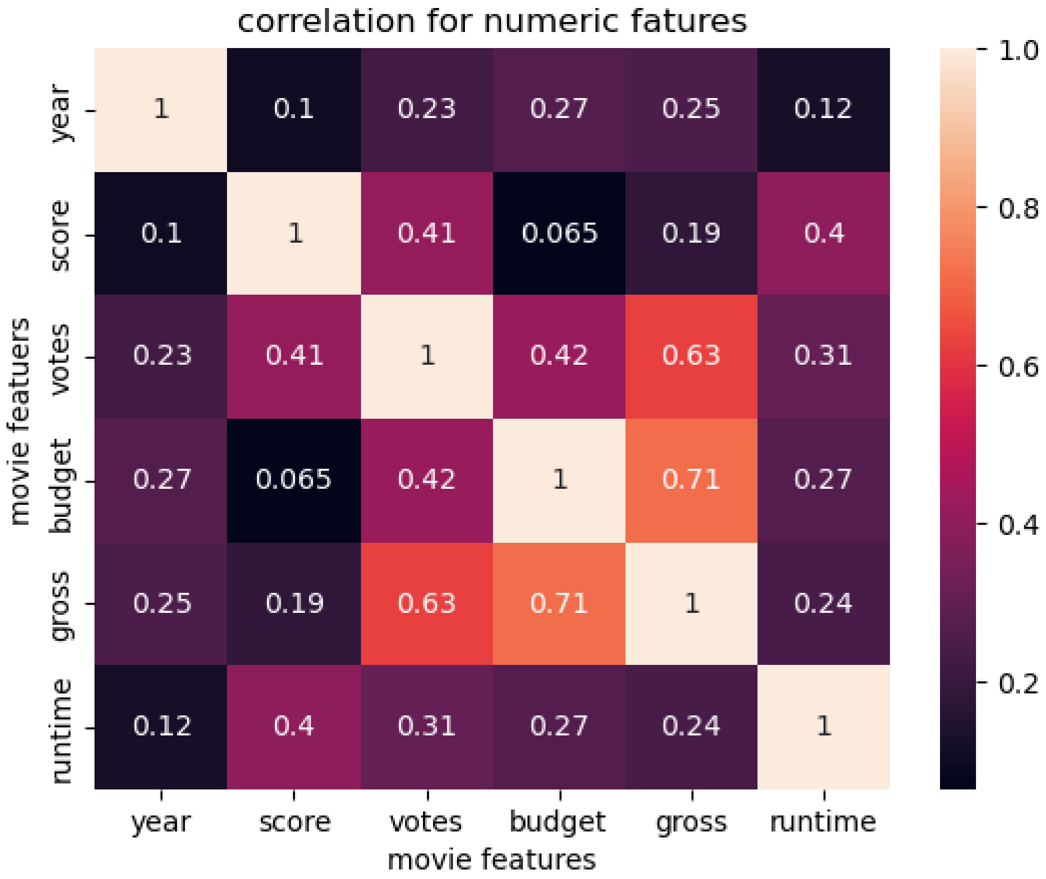
	year	score	votes	budget	gross	runtime
year	1.000000	0.102325	0.226847	0.268851	0.254213	0.120819
score	0.102325	1.000000	0.411931	0.064677	0.185063	0.400560
votes	0.226847	0.411931	1.000000	0.421225	0.629322	0.309355
budget	0.268851	0.064677	0.421225	1.000000	0.712569	0.265934
gross	0.254213	0.185063	0.629322	0.712569	1.000000	0.241619
runtime	0.120819	0.400560	0.309355	0.265934	0.241619	1.000000

In [12]:

```
correlation=df.corr(method='pearson')
sns.heatmap(correlation,annot=True)
plt.title('correlation for numeric fatures')

plt.xlabel('movie features')
plt.ylabel('movie featuers')
plt.show()
```

```
C:\Users\HARDIK\AppData\Local\Temp\ipykernel_3660\2379710088.py:1: FutureW
arning: The default value of numeric_only in DataFrame.corr is deprecated.
In a future version, it will default to False. Select only valid columns o
r specify the value of numeric_only to silence this warning.
  correlation=df.corr(method='pearson')
```



In [13]:

```
df.head(3)
```

Out[13]:

	name	rating	genre	year	released	score	votes	director	write
5445	Avatar	PG-13	Action	2009	December18,2009	7.8	1100000.0	James Cameron	Jame Camero
7445	Avengers: Endgame	PG-13	Action	2019	April26,2019	8.4	903000.0	Anthony Russo	Christophe Marku
3045	Titanic	PG-13	Drama	1997	December19,1997	7.8	1100000.0	James Cameron	Jame Camero

In [14]:

```
df['genre'].unique()
```

Out[14]:

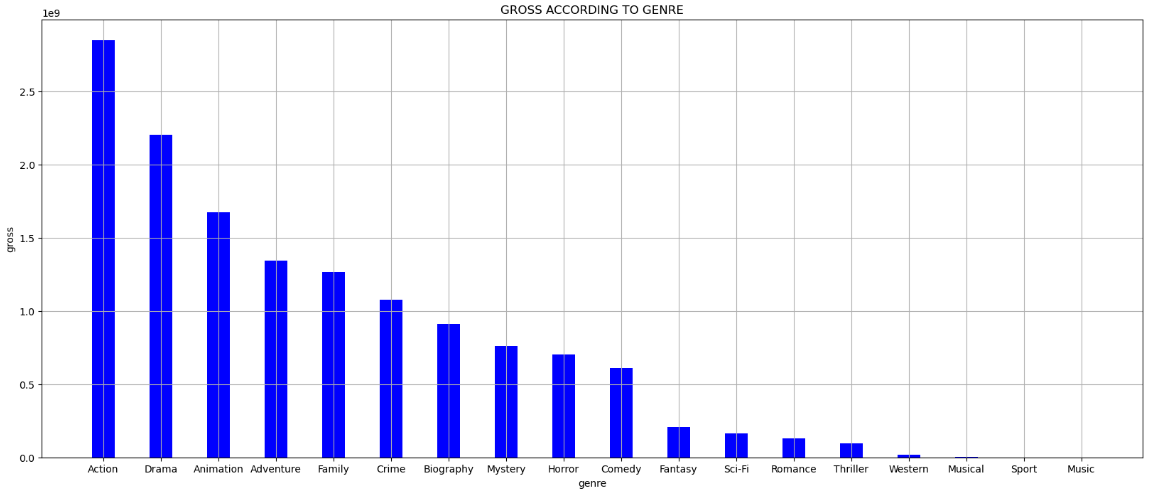
```
array(['Action', 'Drama', 'Animation', 'Adventure', 'Family', 'Crime',
       'Biography', 'Mystery', 'Horror', 'Comedy', 'Fantasy', 'Sci-Fi',
       'Romance', 'Thriller', 'Western', 'Musical', 'Sport', 'Music'],
      dtype=object)
```

In [15]:

```
x=df['genre']
y=df['gross']
plt.figure(figsize=(20,8))
#sns.set(font_scale=1)
plt.grid()
plt.title("GROSS ACCORDING TO GENRE")
plt.xlabel("genre")
plt.ylabel("gross")
plt.bar(x,y,width=0.4,color='b')
```

Out[15]:

```
<BarContainer object of 7575 artists>
```

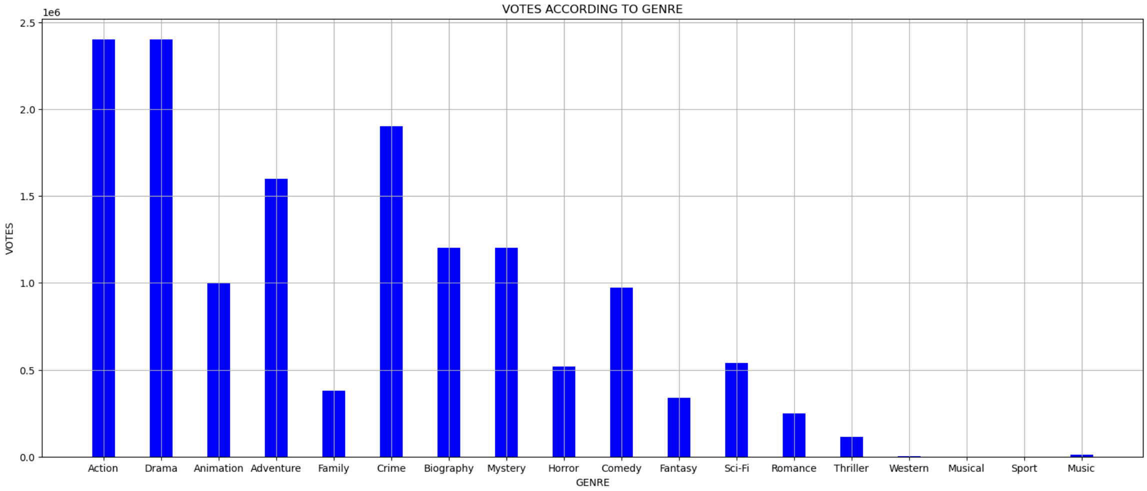


In [16]:

```
x=df['genre']
y=df['votes']
plt.figure(figsize=(20,8))
#sns.set(font_scale=1)
plt.grid()
plt.title("VOTES ACCORDING TO GENRE")
plt.xlabel("GENRE")
plt.ylabel("VOTES")
plt.bar(x,y,width=0.4,color='b')
```

Out[16]:

```
<BarContainer object of 7575 artists>
```

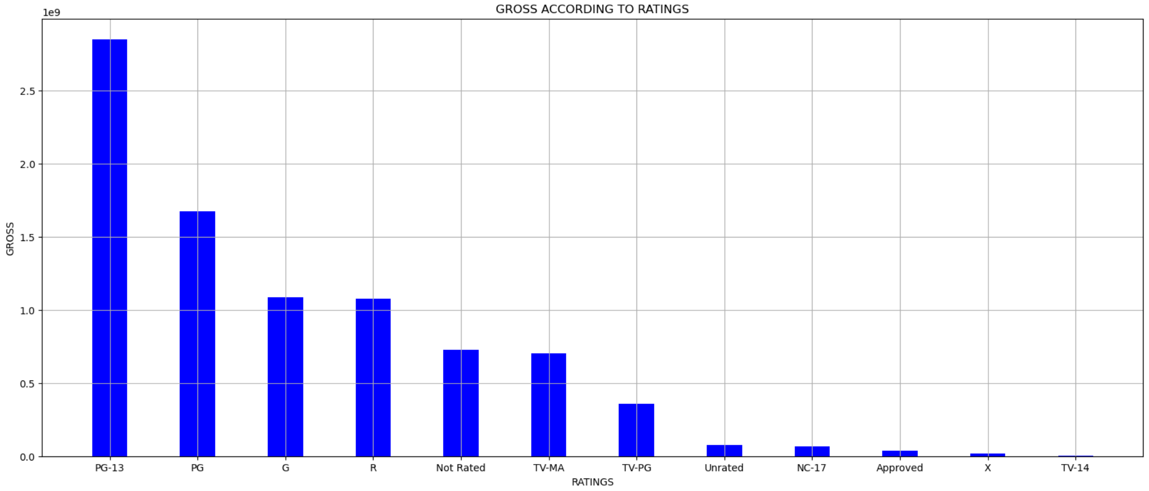


In [17]:

```
x=df['rating']
y=df['gross']
plt.figure(figsize=(20,8))
#sns.set(font_scale=1)
plt.grid()
plt.title("GROSS ACCORDING TO RATINGS")
plt.xlabel("RATINGS")
plt.ylabel("GROSS")
plt.bar(x,y,width=0.4,color='b')
```

Out[17]:

```
<BarContainer object of 7575 artists>
```

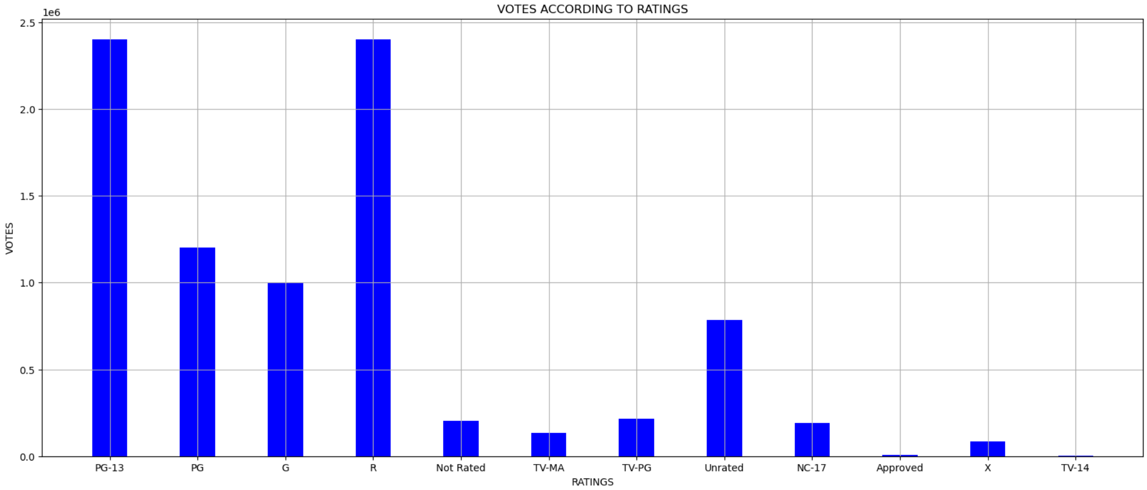


In [18]:

```
x=df['rating']
y=df['votes']
plt.figure(figsize=(20,8))
#sns.set(font_scale=1)
plt.grid()
plt.title("VOTES ACCORDING TO RATINGS")
plt.xlabel("RATINGS")
plt.ylabel("VOTES")
plt.bar(x,y,width=0.4,color='b')
```

Out[18]:

```
<BarContainer object of 7575 artists>
```



In [ ]:

```
```

In [ ]:

```
```