In [1]:
```
import pandas as pd
df=pd.read_csv("C:/Users/Mounika/Downloads/bollywood.csv")
df
```

Out[1]:

	SlNo	Release Date	MovieName	ReleaseTime	Genre	Budget	BoxOfficeCollection	YoutubeViews	YoutubeLikes	YoutubeDislikes
0	1	18-Apr-14	2 States	LW	Romance	36	104.00	8576361	26622	2527
1	2	4-Jan-13	Table No. 21	N	Thriller	10	12.00	1087320	1129	137
2	3	18-Jul-14	Amit Sahni Ki List	N	Comedy	10	4.00	572336	586	54
3	4	4-Jan-13	Rajdhani Express	N	Drama	7	0.35	42626	86	19
4	5	4-Jul-14	Bobby Jasoos	N	Comedy	18	10.80	3113427	4512	1224
...	...	...	...	...	...	...	...	...	...	...
144	145	27-Feb-15	Dum Laga Ke Haisha	N	Comedy	15	30.00	3250917	8185	615
145	146	13-Mar-15	NH10	N	Thriller	13	32.10	5592977	15464	1513
146	147	20-Mar-15	Dilliwali Zaalim Girlfriend	N	Comedy	32	12.00	2316047	4289	807
147	148	20-Mar-15	Hunterrr	N	Comedy	5	11.89	4674795	3706	762
148	149	23-May-14	Kochadaiiyaan	HS	Action	150	120.00	4740727	13466	2649

149 rows × 10 columns

In [2]:
```
print(df.head(10))
```

```
   SlNo Release Date                    MovieName ReleaseTime      Genre  \
0     1  18-Apr-14                      2 States           LW    Romance
1     2   4-Jan-13                  Table No. 21            N   Thriller
2     3  18-Jul-14           Amit Sahni Ki List            N     Comedy
3     4   4-Jan-13               Rajdhani Express            N      Drama
4     5   4-Jul-14                  Bobby Jasoos            N     Comedy
5     6  30-May-14                    Citylights           HS      Drama
6     7  19-Sep-14              Daawat-E-Ishq              N     Comedy
7     8  11-Jan-13  Matru Ki Bijlee Ka Mandola              N     Comedy
8     9  10-Jan-14                 Dedh Ishqiya           LW     Comedy
9    10  11-Jan-13                   Gangoobai              N      Drama

   Budget  BoxOfficeCollection  YoutubeViews  YoutubeLikes  YoutubeDislikes
0      36                104.00       8576361         26622             2527
1      10                 12.00       1087320          1129              137
2      10                  4.00        572336           586               54
3       7                  0.35         42626            86               19
4      18                 10.80       3113427          4512             1224
5       7                 35.00       1074591          1806               84
6      30                 24.40       3905050          8315             1373
7      33                 40.00       2435283          4326              647
8      31                 27.00       2330067          2436              591
9       2                  0.01          4354             1                1
```

In [3]:
```
print(df.tail(10))
```

```
     SlNo Release Date                    MovieName ReleaseTime     Genre  \
139   140   30-Jan-15                  Hawaizaada            N     Drama
140   141   30-Jan-15                Khamoshiyan             N  Thriller
141   142    6-Feb-15                  Shamitabh             N     Drama
142   143   13-Feb-15                        Roy            FS   Romance
143   144   20-Feb-15                  Badlapur            FS    Action
144   145   27-Feb-15  Dum Laga Ke Haisha                    N    Comedy
145   146   13-Mar-15                      NH10            N  Thriller
146   147   20-Mar-15  Dilliwali Zaalim Girlfriend           N    Comedy
147   148   20-Mar-15                  Hunterrr              N    Comedy
148   149   23-May-14               Kochadaiiyaan           HS    Action

     Budget  BoxOfficeCollection  YoutubeViews  YoutubeLikes  YoutubeDislikes
139      25                 30.25       2368404          8619              539
140      11                 14.02       3094001          4599              997
141      40                 38.00       2103508          5599              677
142      40                 58.00       7687797         18974             3229
143      23                 77.00       4550051         10402              893
144      15                 30.00       3250917          8185              615
145      13                 32.10       5592977         15464             1513
146      32                 12.00       2316047          4289              807
147       5                 11.89       4674795          3706              762
148     150                120.00       4740727         13466             2649
```

In [34]:
```
#1.no.of records
print(df.shape[0])
#1.meta data information of dataset
df.describe()
```

```
149
```

Out[34]:

	SlNo	Budget	BoxOfficeCollection	YoutubeViews	YoutubeLikes	YoutubeDislikes
count	149.000000	149.000000	149.000000	1.490000e+02	149.000000	149.000000
mean	75.000000	29.442953	55.667248	3.337920e+06	7877.536913	1207.818792
std	43.156691	28.237981	94.494531	3.504407e+06	12748.047191	1852.692938
min	1.000000	2.000000	0.010000	4.354000e+03	1.000000	1.000000
25%	38.000000	11.000000	8.780000	1.076591e+06	1377.000000	189.000000
50%	75.000000	21.000000	28.000000	2.375050e+06	4111.000000	614.000000
75%	112.000000	35.000000	57.450000	4.550051e+06	9100.000000	1419.000000
max	149.000000	150.000000	735.000000	2.317107e+07	101275.000000	11888.000000

In [37]:
```
'''2.How many movies got released in each genre? Which genre had highest number of releases? Sort
number of releases in each genre in descending order.'''
print(df.Genre.value_counts())
print(max(df.Genre.value_counts()))
```

```
Comedy       36
 Drama       35
Thriller     26
Romance      25
Action       21
Action        3
Thriller      3
Name: Genre, dtype: int64
36
```

In [39]:
```
'''3.How many movies in each genre got released in different release times like long weekend, festive
season, etc.'''
pd.crosstab(df.Genre,df.ReleaseTime)
```

Out[39]:

ReleaseTime	FS	HS	LW	N
Genre				
Drama	4	6	1	24
Action	3	3	3	12
Action	0	0	0	3
Comedy	3	5	5	23
Romance	3	3	4	15
Thriller	4	1	1	20
Thriller	0	0	1	2

In [46]:
```
'''4.Which month of the year, maximum number movie releases are seen? '''
df['Release Date']=pd.to_datetime(df['Release Date'])
df["Year"]=df["Release Date"].dt.year
print(df.Year.value_counts())
print(max(df.Year.value_counts()))
#maximum no.of movies are released in 2014
```

```
2014    70
2013    67
2015    12
Name: Year, dtype: int64
70
```

In [48]:
```
'''5.Which month of the year typically sees most releases of high budgeted movies, that is, movies with
budget of 25 crore or more?'''
small_df=df[df.Budget>=10]
print(small_df['Release Date'].dt.month.value_counts())
#month2-february sees most releases of high budgeted movies
```

```
2     8
8     7
11    6
1     6
7     5
6     5
10    4
9     4
5     3
4     3
3     3
12    2
Name: Release Date, dtype: int64
```

In [49]:
```
'''6.Which are the top 10 movies with maximum return on investment (ROI)? '''
df['ROI']=(df.BoxOfficeCollection-df.Budget)/df.Budget
df.sort_values(by='ROI').MovieName[0:10]
```

Out[49]:

```
9                Gangoobai
15                Bandook
53                Sona Spa
3          Rajdhani Express
49     Kya Dilli Kya Lahore
121               Satya 2
67             Purani Jeans
103         Samrat and Co.
30               Heartless
102               Kaanchi
Name: MovieName, dtype: object
```

In [50]:
```
'''7.Do the movies have higher ROI if they get released on festive seasons or long weekend? Calculate
the average ROI for different release times'''
df.groupby(by='ReleaseTime').ROI.mean()
#movies release on lw and fs have higher values of ROI-1.12 and 0.97
```

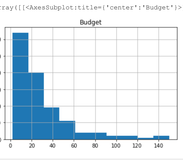
Out[50]:

```
ReleaseTime
FS    0.973853
HS    0.850867
LW    1.127205
N     0.657722
Name: ROI, dtype: float64
```

In [51]:
```
'''8.Draw a histogram and a distribution plot to find out the distribution of movie budgets. Interpret the
plot to conclude if the most movies are high or low budgeted movies.'''
import matplotlib.pyplot as plt
df.hist('Budget')
```

Out[51]:

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



In [55]:
```
'''9.Compare the distribution of ROIs between movies with comedy genre and drama. Which genre
typically sees higher ROIs?'''
df['ROI']=(df.BoxOfficeCollection-df.Budget)/df.Budget
print(df['ROI'].MovieName.value_counts())
print(df['ROI'].Genre.value_counts())
```

```
---------------------------------------------------------------------------
AttributeError                            Traceback (most recent call last)
<ipython-input-55-b9a8cbad02fd> in <module>
      2 typically sees higher ROIs?'''
      3 df['ROI']=(df.BoxOfficeCollection-df.Budget)/df.Budget
----> 4 print(df['ROI'].MovieName.value_counts())
      5 print(df['ROI'].Genre.value_counts())

~\anaconda\lib\site-packages\pandas\core\generic.py in __getattr__(self, name)
   5137             if self._info_axis._can_hold_identifiers_and_holds_name(name):
   5138                 return self[name]
-> 5139             return object.__getattribute__(self, name)
   5140
   5141     def __setattr__(self, name: str, value) -> None:

AttributeError: 'Series' object has no attribute 'MovieName'
```

In [59]:
```
'''10.Is there a correlation between box office collection and YouTube likes? Is the correlation positive or
negative?'''
corr=df[['BoxOfficeCollection','YoutubeLikes']].corr()
corr
#boc increases 100% with correlation coefficient 1.00 and youtubelikes increases 68% with corr coef 0.68
```

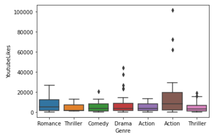
Out[59]:

	BoxOfficeCollection	YoutubeLikes
BoxOfficeCollection	1.000000	0.682517
YoutubeLikes	0.682517	1.000000

In [61]:
```
'''11.Which genre of movies typically sees more YouTube likes? Draw boxplots for each genre of movies
to compare'''
import seaborn as sns
sns.boxplot(x='Genre',y='YoutubeLikes',data=df)
#genre action has highest youtube likes
```

Out[63]:

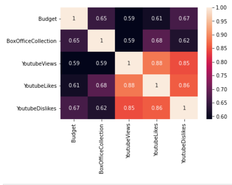
```
<AxesSubplot:xlabel='Genre', ylabel='YoutubeLikes'>
```



In [64]:
```
'''12.Which of the variables among Budget, BoxOfficeCollection, YoutubeView, YoutubeLikes,
YoutubeDislikes are highly correlated? Note: Draw pair plot or heatmap.'''
corr=df[['Budget','BoxOfficeCollection','YoutubeViews','YoutubeLikes','YoutubeDislikes']].corr()
sns.heatmap(corr,annot=True)
```

Out[64]:

```
<AxesSubplot:>
```



In [ ]:
```
```