Krittamate Cherdpongtagit 6310545884

Data Exploration Exercise

1. How many unique genres are there ?
   Answer : 13

```
In [7]: df1 = df

In [8]: selected_var = "Genre"

In [9]: df1[selected_var].describe()

Out[9]: count       1000
        unique        13
        top       Action
        freq         293
        Name: Genre, dtype: object
```

---

2. Which director produced the longest movie?
   Answer : Robert Rodriguez

```
In [25]: df1.groupby('Runtime').max()
```

Out[25]:

Runtime	Title	Genre	Director	Year	Rating	Votes
66	Ma vie de Courgette	Animation	Claude Barras	2016	7.8	4370
73	Wolves at the Door	Horror	Shawn Burkett	2016	4.6	564
80	La tortue rouge	Animation	Michael Dudok de Wit	2016	7.6	11482
81	The Thinning	Thriller	Mike Flanagan	2016	6.6	69823
82	Tramps	Comedy	Adam Leon	2016	6.5	1031
...	...	...	...	...	...	...
170	3 Idiots	Comedy	Rajkumar Hirani	2009	8.4	238789
172	Cloud Atlas	Drama	Tom Tykwer	2012	7.5	298651
180	The Wolf of Wall Street	Drama	Martin Scorsese	2013	8.2	865134
187	The Hateful Eight	Crime	Quentin Tarantino	2015	7.8	341170
191	Grindhouse	Action	Robert Rodriguez	2007	7.6	160350

94 rows × 6 columns

3. What is the average runtime of horror movies?
   Answer : 97.76 minutes

```
In [29]: df1.groupby(df1.Genre == 'Horror').mean()
```

Out[29]:

Genre	Year	Runtime	Rating	Votes
False	2012.759958	113.915094	6.764465	174165.855346
True	2013.260870	97.760870	5.867391	79435.413043

---

4. Which year did the movie have the highest average rating?
   Answer : 2007

```
In [30]: df1.groupby("Year").mean()
```

Out[30]:

Year	Runtime	Rating	Votes
2006	120.840909	7.125000	269289.954545
2007	121.622642	7.133962	244331.037736
2008	110.826923	6.784615	275505.384615
2009	116.117647	6.960784	255780.647059
2010	111.133333	6.826667	252782.316667
2011	114.603175	6.838095	240790.301587
2012	119.109375	6.925000	285226.093750
2013	116.065934	6.812088	219049.648352
2014	114.489796	6.837755	203930.224490
2015	114.496063	6.602362	115726.220472
2016	107.373737	6.436700	48591.754209

---

5. What was the latest movie from the director who directed the most movies?
   Answer: Ridley Scott

```
In [31]: df1["Director"].describe()

Out[31]: count             1000
         unique             644
         top       Ridley Scott
         freq                 8
         Name: Director, dtype: object
```

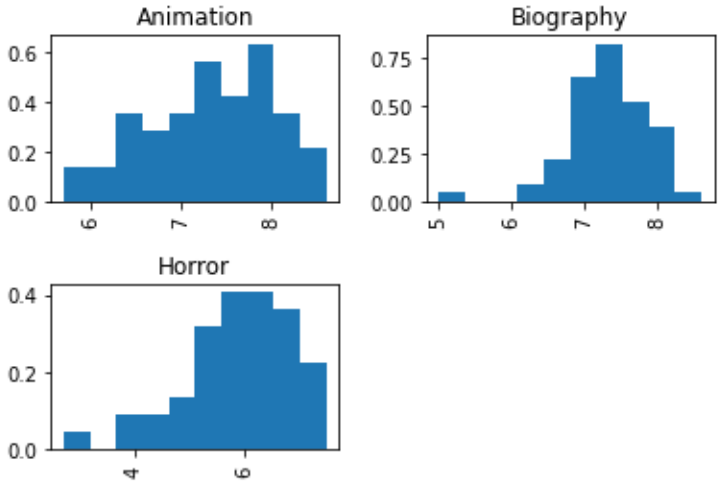
---

6. Plot histogram of rating probabilities for the following 3 genres: animation, biography and horror. One graph per one genre. Describe the comparison of rating distributions between these 3 genres.
   Answer :

- Horror genre has a lower central tendency than the other two.
- Three histograms have similar bell shaped distribution.
- Horror genre has the lowest outlier.
- Biography genre has the highest probability of rating at about 0.8

```
In [28]: selected_var = 'Rating'
         temp_df = df[df.Genre.isin(['Animation','Biography','Horror'])]
         temp_df.hist(selected_var, by='Genre', density=True)
         plt.title('Probability of Rating')
         plt.ylabel('Probability')

Out[28]: Text(0, 0.5, 'Probability')
```



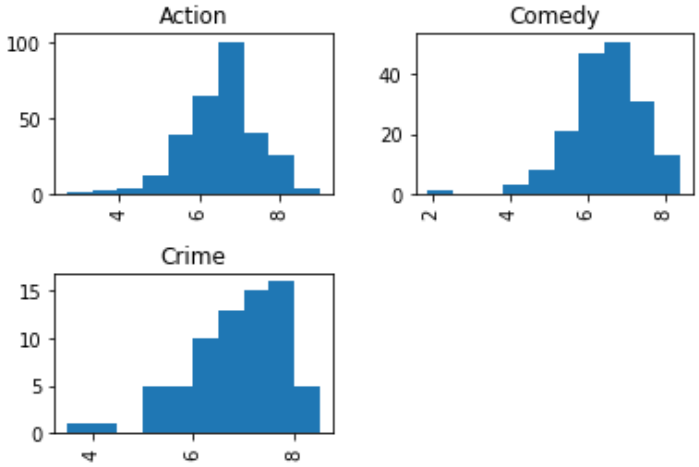
7. Plot an overlay histogram of rating frequencies for the following 3 genres: action, comedy, and crime. What can you tell about movies from these 3 genres based on the overlay histogram ?

   Answer :

- Action genre has the highest frequency of rating, crime has the lowest.
- Three histograms have similar bell shaped distribution.

```
In [29]: selected_var = 'Rating'
         temp_df = df[df.Genre.isin(['Action','Comedy','Crime'])]
         temp_df.hist(selected_var, by='Genre')
         plt.title('Frequency of Rating')
         plt.ylabel('Frequency')

Out[29]: Text(0, 0.5, 'Frequency')
```



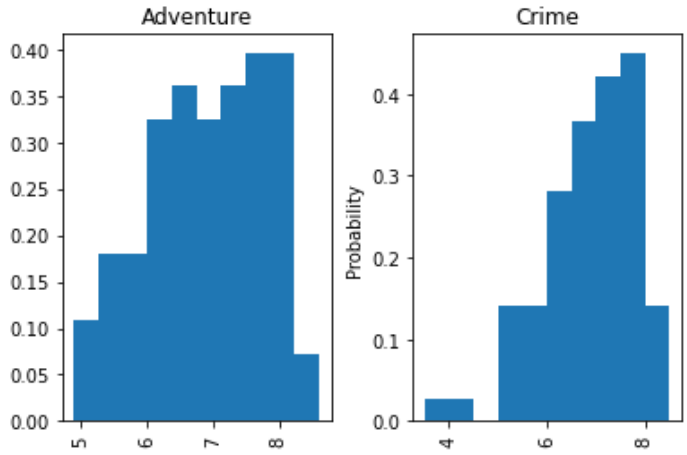
---

8. Was adventure or crime movie likely to receive a higher rating?

Answer : Crime movie

```
In [34]: selected_var = 'Rating'
         temp_df = df[df.Genre.isin(['Adventure','Crime'])]
         temp_df.hist(selected_var, by='Genre', density=True)
         plt.ylabel('Probability')

Out[34]: Text(0, 0.5, 'Probability')
```

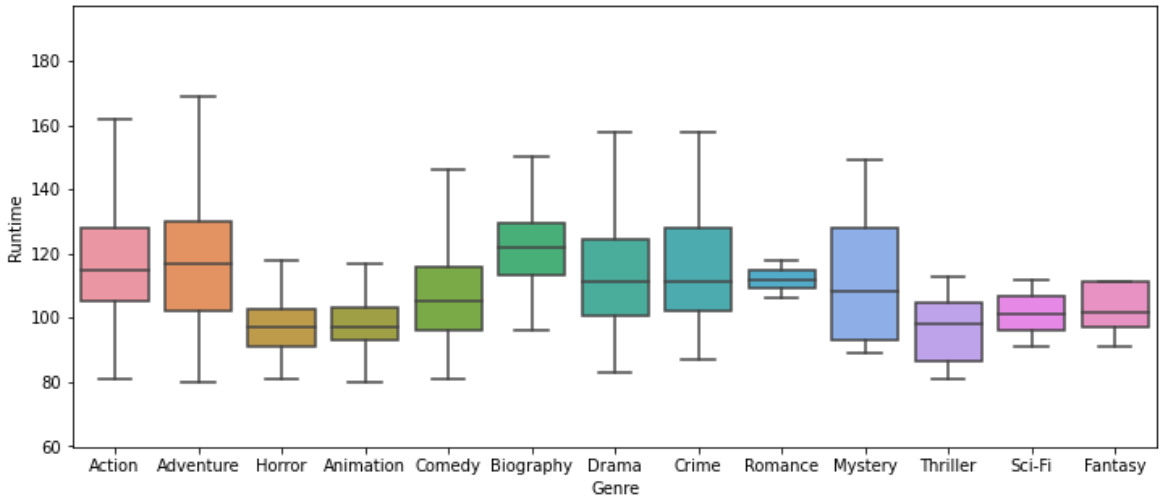


9. Plot a boxplot of runtime for all genres. Which genre tends to have the shortest and the longest runtime? Provide the reason.

   Answer : Thriller genre tends to have the shortest runtime, adventure genre tends to have the longest. Because the boxplot of thriller genre has the lowest central tendency of runtime and adventure has the highest.

In [47]:
```
plt.figure(figsize=(12,5))
sns.boxplot(x='Genre', y='Runtime', data=df, fliersize=0)
```

Out[47]: <AxesSubplot:xlabel='Genre', ylabel='Runtime'>



---

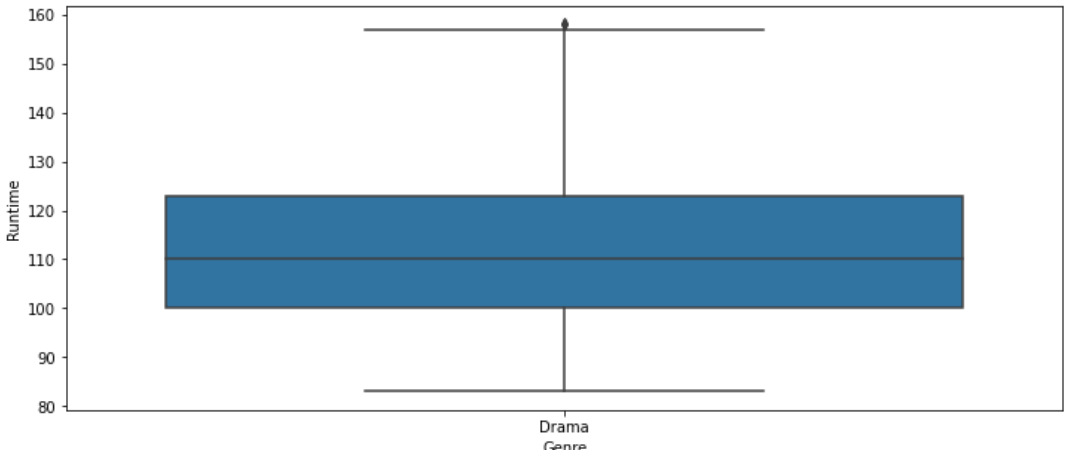
10. Plot 3 boxplots of runtime for the drama genre. One for each data set after outliers are removed by (1) IQR-based, (2) SD-based, and (3) 5% and 95% based. Do these 3 boxplots look the same or not? Why did they look the same or different ?

Answer : They look a bit different because the outlier are removed in different way

1) Removed by IQR-based

In [61]:
```
outlier_var = 'Runtime'
df2 = df
df2 = df2[df2.Genre == 'Drama']
q1 = df2[outlier_var].quantile(0.25)
q3 = df2[outlier_var].quantile(0.75)
iqr = q3-q1
temp_df = df2[~((df2[outlier_var] < q1-1.5*iqr) | (df2[outlier_var] > q3+1.5*iqr))]
plt.figure(figsize=(12,5))
sns.boxplot(x='Genre', y='Runtime', data=temp_df)
```

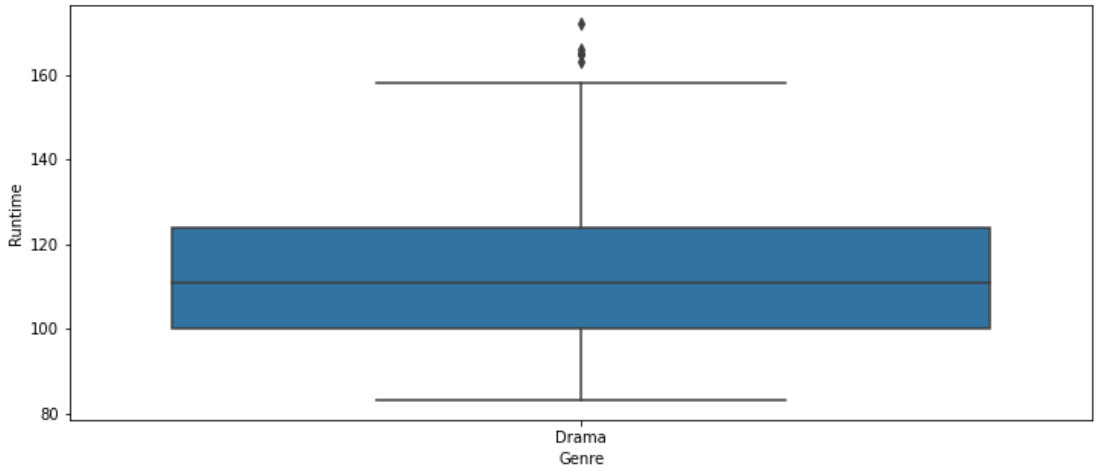
Out[61]: <AxesSubplot:xlabel='Genre', ylabel='Runtime'>



2) Removed by SD-based

In [62]:
```
mean = df2[outlier_var].mean()
sd = df2[outlier_var].std()
temp_df2 = df2[~((df2[outlier_var] < mean-3*sd) | (df2[outlier_var] > mean+3*sd))]
plt.figure(figsize=(12,5))
sns.boxplot(x='Genre', y='Runtime', data=temp_df2)
```

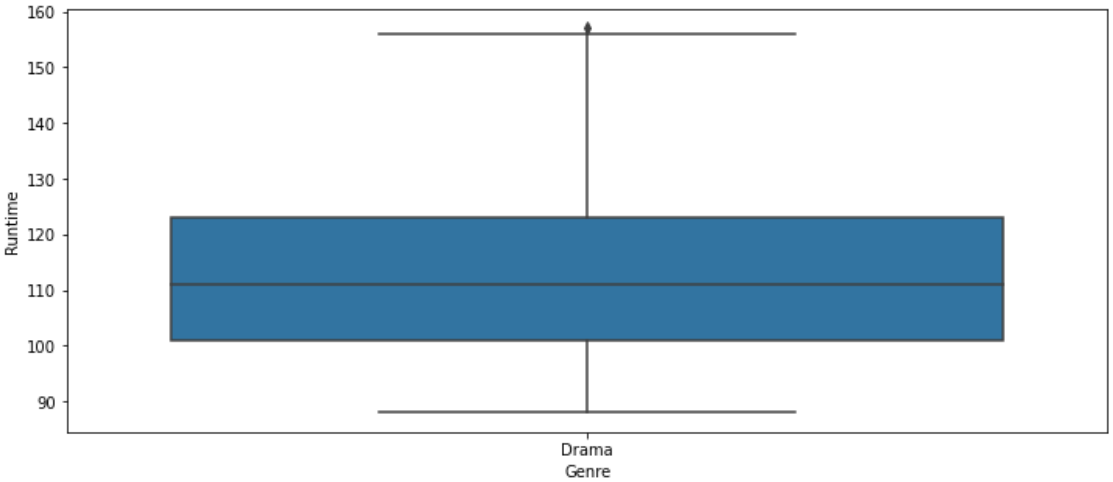
Out[62]: <AxesSubplot:xlabel='Genre', ylabel='Runtime'>



3) Removed by 5% and 95% based

In [63]:
```
p05 = df2[outlier_var].quantile(0.05)
p95 = df2[outlier_var].quantile(0.95)
temp_df3 = df2[~((df2[outlier_var] < p05) | (df2[outlier_var] > p95))]
plt.figure(figsize=(12,5))
sns.boxplot(x='Genre', y='Runtime', data=temp_df3)
```

Out[63]: <AxesSubplot:xlabel='Genre', ylabel='Runtime'>



11. In the year with the maximum movie production, create a graph to show the percentage of each movie genre. (Do not provide the numerical answer only.)
    Answer:

Find the year with maximum movie production

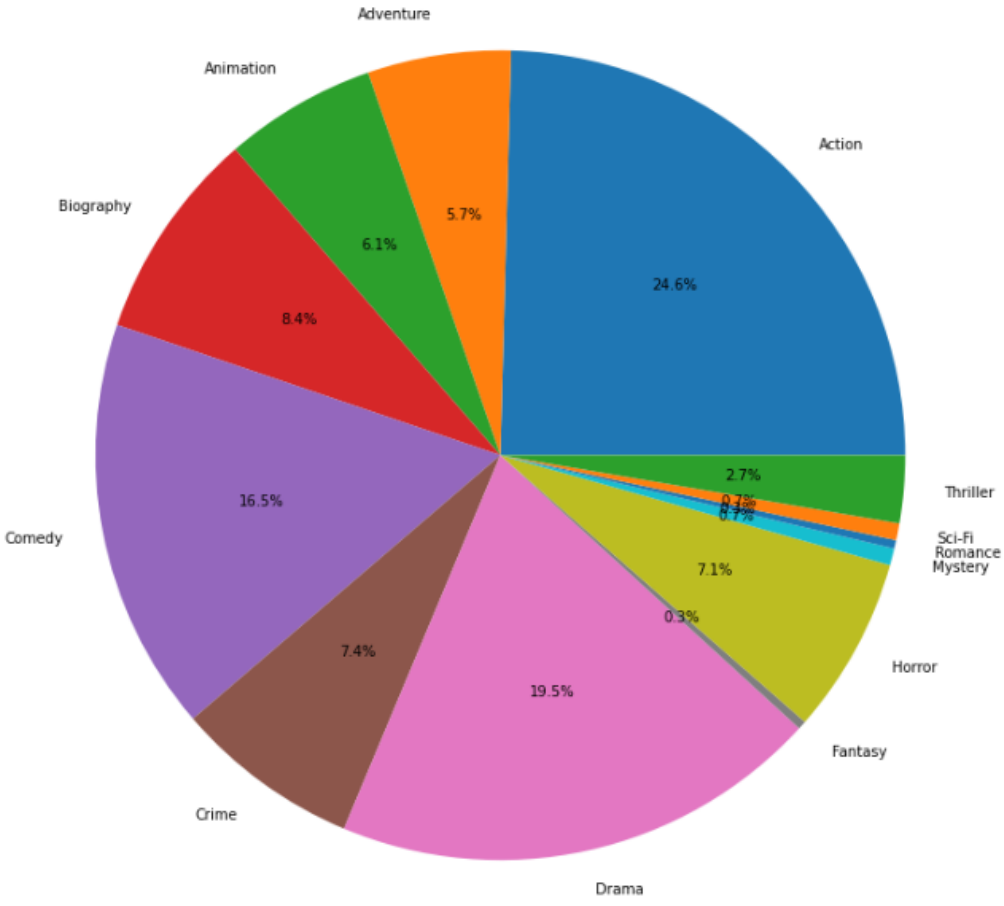
```
In [22]: temp_df = df
         temp_df.groupby("Year").count()
```

Out[22]:

Year	Title	Genre	Director	Runtime	Rating	Votes
2006	44	44	44	44	44	44
2007	53	53	53	53	53	53
2008	52	52	52	52	52	52
2009	51	51	51	51	51	51
2010	60	60	60	60	60	60
2011	63	63	63	63	63	63
2012	64	64	64	64	64	64
2013	91	91	91	91	91	91
2014	98	98	98	98	98	98
2015	127	127	127	127	127	127
2016	297	297	297	297	297	297

Plot the graph

```
In [55]: df3 = df[df.Year == 2016]
         temp_df = df3.groupby("Genre").count()
         fig = plt.figure(figsize=(13,13))
         plt.pie(temp_df['Title'], labels = temp_df.index, autopct='%1.1f%%')
         plt.savefig('survived_embarked_pie.jpg')
```



12. How many series of X-men movies are there ? In addition, find out which year each series was produced?
    Answer: There're 4 X-men movies produced in 2006, 2009, 2014, 2016

```
In [76]: temp_df = df[df['Title'].str.contains("X-Men")]
         temp_df
```

Out[76]:

	Title	Genre	Director	Year	Runtime	Rating	Votes
32	X-Men: Apocalypse	Action	Bryan Singer	2016	144	7.1	275510
162	X-Men: Days of Future Past	Action	Bryan Singer	2014	132	8.0	552298
268	X-Men Origins: Wolverine	Action	Gavin Hood	2009	107	6.7	388447
626	X-Men: The Last Stand	Action	Brett Ratner	2006	104	6.7	406540

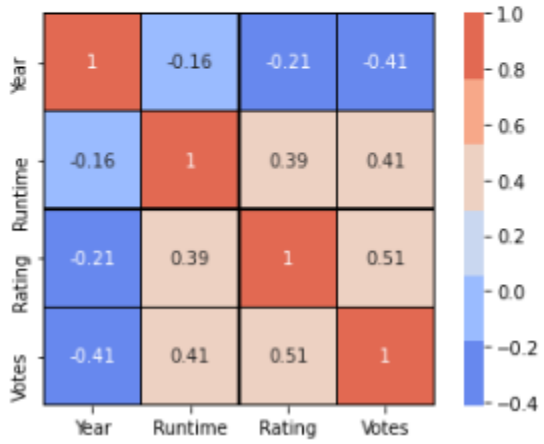
---

13. Overall, what correlates with rating more? Runtime or Votes? Did correlation of rating and runtime and correlation of rating and votes change over time? Show how you justify your answer.

Answer: Votes correlates with rating more.
The correlation of rating and runtime and correlation of rating and votes don't change over time.
Because the correlation between votes and rating is more than runtime and rating and the value don't change
(in (Rating,Votes)=0.51 in (Votes,Rating) =0.51)
(in (Rating,Runtime)=0.39 in (Runtime,Rating) =0.39)

```
In [98]: sns.heatmap(temp_df.corr(),
                     square=True,
                     linewidths=0.25,
                     linecolor=(0,0,0),
                     cmap=sns.color_palette("coolwarm"),
                     annot=True)
```

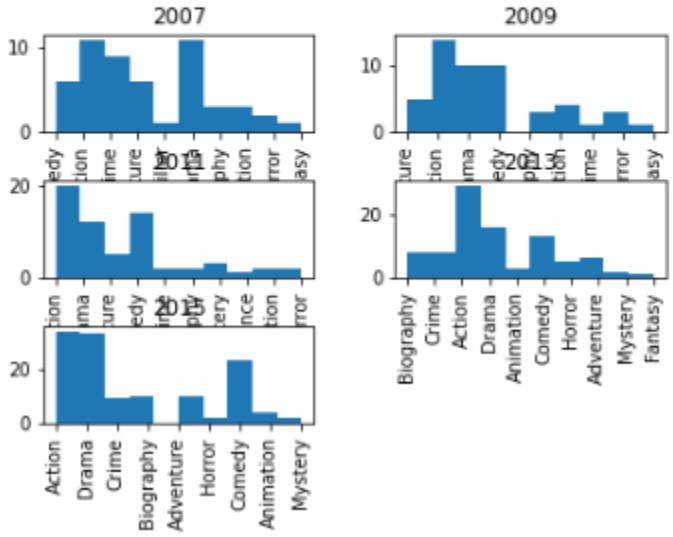
```
Out[98]: <matplotlib.axes._subplots.AxesSubplot at 0x1fdf20d2ac0>
```



14. For each year that is even number, what are the top 3 genres with maximum number of movies ?

```
In [192]: selected_var = 'Genre'
          temp_df = df[(df.Year % 2) != 0]
          temp_df.hist(selected_var, by='Year')
          plt.title('Frequency of movies')
          plt.ylabel('Frequency')
```

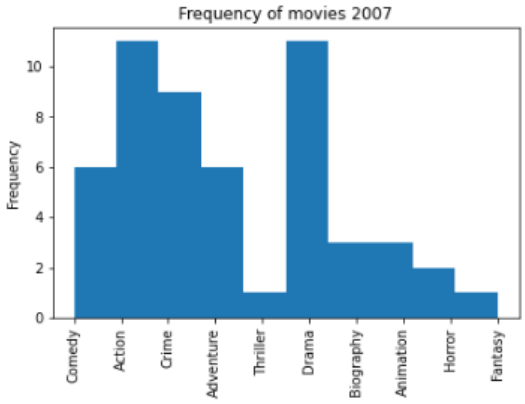
```
Out[192]: Text(0, 0.5, 'Frequency')
```



Answer: 2007 Action:11 Drama:11 Crime:9

```
In [221]: selected_var = 'Genre'
          temp_df = df[df.Year ==  2007]
          temp_df.hist(selected_var, by='Year')
          plt.title('Frequency of movies 2007')
          plt.ylabel('Frequency')
```

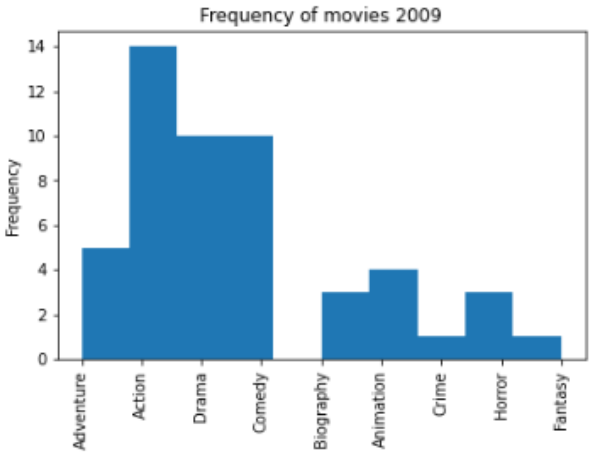
```
Out[221]: Text(0, 0.5, 'Frequency')
```



Answer: 2009 Action:14 Drama:10 Comedy:10

```
In [222]: selected_var = 'Genre'
          temp_df = df[df.Year ==  2009]
          temp_df.hist(selected_var, by='Year')
          plt.title('Frequency of movies 2009')
          plt.ylabel('Frequency')
```

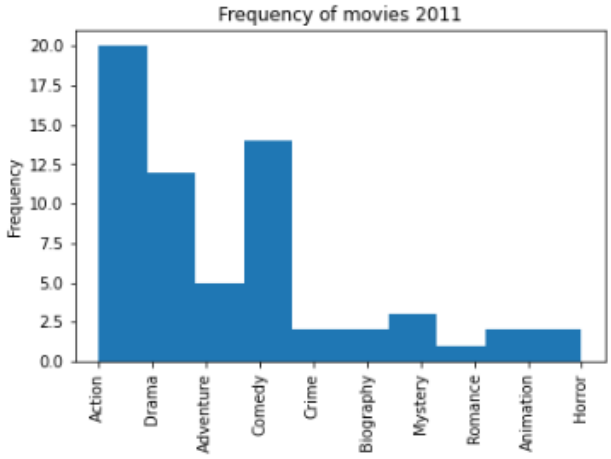
```
Out[222]: Text(0, 0.5, 'Frequency')
```



Answer: 2011 Action:20 Comedy:14 Drama:12

```
In [223]: selected_var = 'Genre'
          temp_df = df[df.Year ==  2011]
          temp_df.hist(selected_var, by='Year')
          plt.title('Frequency of movies 2011')
          plt.ylabel('Frequency')
```

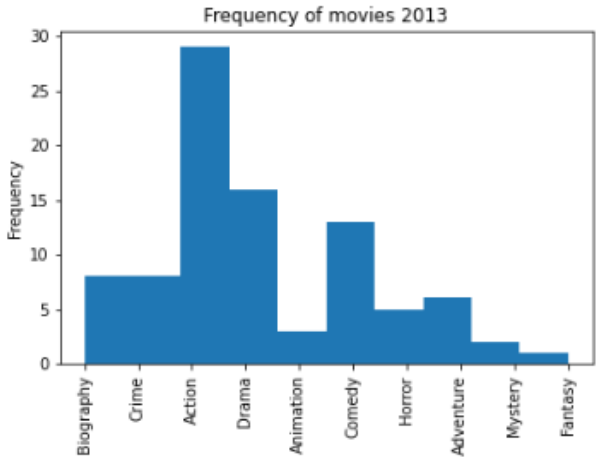
```
Out[223]: Text(0, 0.5, 'Frequency')
```



Answer: 2013 Action:28 Drama:16 Comedy:13

```
In [224]: selected_var = 'Genre'
          temp_df = df[df.Year ==  2013]
          temp_df.hist(selected_var, by='Year')
          plt.title('Frequency of movies 2013')
          plt.ylabel('Frequency')
```

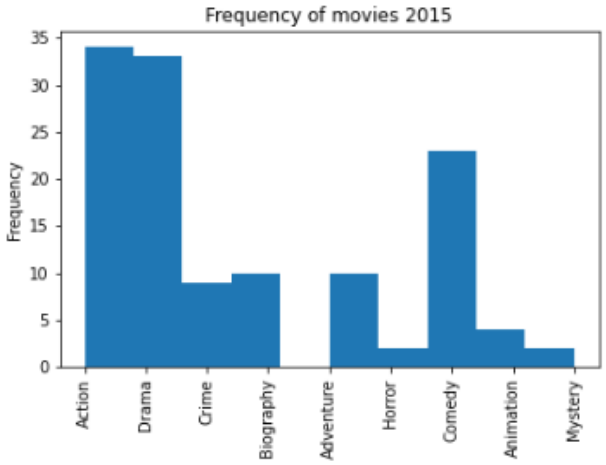
```
Out[224]: Text(0, 0.5, 'Frequency')
```



Answer: 2015 Action:34 Drama:33 Comedy:22

```
In [225]: selected_var = 'Genre'
          temp_df = df[df.Year ==  2015]
          temp_df.hist(selected_var, by='Year')
          plt.title('Frequency of movies 2015')
          plt.ylabel('Frequency')
```

```
Out[225]: Text(0, 0.5, 'Frequency')
```



15. Create a new class of movies by rating, where ‘S-class’ rating = 8.00-10.00, ‘A-class’ rating = 5.00-8.00, and ‘B-class’ rating = 0.00-4.99. Create a graph to show how many movies are in each class. (Do not provide the numerical answer only.)

In [237]:
```
df4 = df.sort_values(by='Rating',ascending=False,inplace=False)
df4['Class'] = pd.cut(df4['Rating'], bins=[0, 4.99, 8, 10], labels=['B-class', 'A-class', 'S-class'])
df4['Class'].value_counts().plot(kind='bar')
plt.xlabel('Movie Class')
plt.ylabel('Number of Movies')
plt.title('Number of Movies in Each Class')
plt.show()
```