```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
from sklearn.model_selection import train_test_split
from sklearn.svm import SVC
```

```
dat = pd.read_csv("/content/IMDb Movies India.csv", encoding='latin-1')
```

```
dat
```

	Name	Year	Duration	Genre	Rating	Votes	Director	Actor 1	Actor 2	Actor 3
0		NaN	NaN	Drama	NaN	NaN	J.S. Randhawa	Manmauji	Birbal	Rajendra Bhatia
1	#Gadhvi (He thought he was Gandhi)	(2019)	109 min	Drama	7.0	8	Gaurav Bakshi	Rasika Dugal	Vivek Ghamande	Arvind Jangid
2	#Homecoming	(2021)	90 min	Drama, Musical	NaN	NaN	Soumyajit Majumdar	Sayani Gupta	Plabita Borthakur	Roy Angana
3	#Yaaram	(2019)	110 min	Comedy, Romance	4.4	35	Ovais Khan	Prateik	Ishita Raj	Siddhant Kapoor
4	...And Once Again	(2010)	105 min	Drama	NaN	NaN	Amol Palekar	Rajat Kapoor	Rituparna Sengupta	Antara Mali
...	...	...	...	...	...	...	...	...	...	...
15504	Zulm Ko Jala Doonga	(1988)	NaN	Action	4.6	11	Mahendra Shah	Naseeruddin Shah	Sumeet Saigal	Suparna Anand
15505	Zulmi	(1999)	129 min	Action, Drama	4.5	655	Kuku Kohli	Akshay Kumar	Twinkle Khanna	Aruna Irani
15506	Zulmi Raj	(2005)	NaN	Action	NaN	NaN	Kiran Thej	Sangeeta Tiwari	NaN	NaN
15507	Zulmi Shikari	(1988)	NaN	Action	NaN	NaN	NaN	NaN	NaN	NaN
15508	Zulm-O-Sitam	(1998)	130 min	Action, Drama	6.2	20	K.C. Bokadia	Dharmendra	Jaya Prada	Arjun Sarja

15509 rows × 10 columns

```
dat.head()
```

	Name	Year	Duration	Genre	Rating	Votes	Director	Actor 1	Actor 2	Actor 3
0		NaN	NaN	Drama	NaN	NaN	J.S. Randhawa	Manmauji	Birbal	Rajendra Bhatia
1	#Gadhvi (He thought he was Gandhi)	(2019)	109 min	Drama	7.0	8	Gaurav Bakshi	Rasika Dugal	Vivek Ghamande	Arvind Jangid
2	#Homecoming	(2021)	90 min	Drama, Musical	NaN	NaN	Soumyajit Majumdar	Sayani Gupta	Plabita Borthakur	Roy Angana
3	#Yaaram	(2019)	110 min	Comedy, Romance	4.4	35	Ovais Khan	Prateik	Ishita Raj	Siddhant Kapoor
4	...And Once Again	(2010)	105 min	Drama	NaN	NaN	Amol Palekar	Rajat Kapoor	Rituparna Sengupta	Antara Mali

```
dat.head(2)
```

	Name	Year	Duration	Genre	Rating	Votes	Director	Actor 1	Actor 2	Actor 3
0		NaN	NaN	Drama	NaN	NaN	J.S. Randhawa	Manmauji	Birbal	Rajendra Bhatia
1	#Gadhvi (He thought he was Gandhi)	(2019)	109 min	Drama	7.0	8	Gaurav Bakshi	Rasika Dugal	Vivek Ghamande	Arvind Jangid

```
dat.shape
```

```
(15509, 10)
```

```
dat.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 15509 entries, 0 to 15508
Data columns (total 10 columns):
 #   Column    Non-Null Count  Dtype
---  ------    --------------  -----
 0   Name      15509 non-null  object
 1   Year      14981 non-null  object
 2   Duration  7240 non-null   object
 3   Genre     13632 non-null  object
 4   Rating    7919 non-null   float64
 5   Votes     7920 non-null   object
 6   Director  14984 non-null  object
 7   Actor 1   13892 non-null  object
 8   Actor 2   13125 non-null  object
 9   Actor 3   12365 non-null  object
dtypes: float64(1), object(9)
memory usage: 1.2+ MB
```

```
dat.isnull()
```

	Name	Year	Duration	Genre	Rating	Votes	Director	Actor 1	Actor 2	Actor 3
0	False	True	True	False	True	True	False	False	False	False
1	False	False	False	False	False	False	False	False	False	False
2	False	False	False	False	True	True	False	False	False	False
3	False	False	False	False	False	False	False	False	False	False
4	False	False	False	False	True	True	False	False	False	False
...	...	...	...	...	...	...	...	...	...	...
15504	False	False	True	False	False	False	False	False	False	False
15505	False	False	False	False	False	False	False	False	False	False
15506	False	False	True	False	True	True	False	False	True	True
15507	False	False	True	False	True	True	True	True	True	True
15508	False	False	False	False	False	False	False	False	False	False

15509 rows × 10 columns

```
dat.isnull().sum()
```

	0
Name	0
Year	528
Duration	8269
Genre	1877
Rating	7590
Votes	7589
Director	525
Actor 1	1617
Actor 2	2384
Actor 3	3144

dtype: int64

```
dat.isnull().sum() / dat.shape[0] * 100
```

	0
Name	0.000000
Year	3.404475
Duration	53.317429
Genre	12.102650
Rating	48.939326
Votes	48.932878
Director	3.385131
Actor 1	10.426204
Actor 2	15.371720
Actor 3	20.272100

dtype: float64

```
from sklearn.preprocessing import LabelEncoder

le = LabelEncoder()


dat["Name"] = le.fit_transform(dat["Name"])
dat["Year"] = le.fit_transform(dat["Year"])
dat["Duration"] = le.fit_transform(dat["Duration"])
dat["Genre"] = le.fit_transform(dat["Genre"])
dat["Votes"] = le.fit_transform(dat["Votes"])
dat["Director"] = le.fit_transform(dat["Director"])
dat["Actor 1"] = le.fit_transform(dat["Actor 1"])
```

```
dat["Actor 2"] = le.fit_transform(dat["Actor 2"])
dat["Actor 3"] = le.fit_transform(dat["Actor 3"])
```

```
dat.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 15509 entries, 0 to 15508
Data columns (total 10 columns):
 #   Column    Non-Null Count  Dtype
---  ------    --------------  -----
 0   Name      15509 non-null  int64
 1   Year      15509 non-null  int64
 2   Duration  15509 non-null  int64
 3   Genre     15509 non-null  int64
 4   Rating    7919 non-null   float64
 5   Votes     15509 non-null  int64
 6   Director  15509 non-null  int64
 7   Actor 1   15509 non-null  int64
 8   Actor 2   15509 non-null  int64
 9   Actor 3   15509 non-null  int64
dtypes: float64(1), int64(9)
memory usage: 1.2 MB
```

```
dat.select_dtypes("int")
```

	Name	Year	Duration	Genre	Votes	Director	Actor 1	Actor 2	Actor 3
0	0	102	182	299	2034	1926	2250	800	3108
1	1	98	9	299	1849	1548	3280	4790	527
2	2	100	172	351	2034	5123	3713	2866	3450
3	3	98	10	228	1169	3319	2917	1504	4020
4	7	89	5	299	2034	385	3112	3462	405
...	...	...	...	...	...	...	...	...	...
15504	13832	67	182	0	368	2690	2586	4299	4262
15505	13834	78	29	40	1687	2499	227	4532	519
15506	13835	84	182	0	2034	2424	3609	4891	4820
15507	13836	67	182	0	2034	5938	4718	4891	4820
15508	13833	77	30	40	794	2195	1139	1589	490

15509 rows × 9 columns

```
from sklearn.preprocessing import OneHotEncoder

ohe = OneHotEncoder()
```

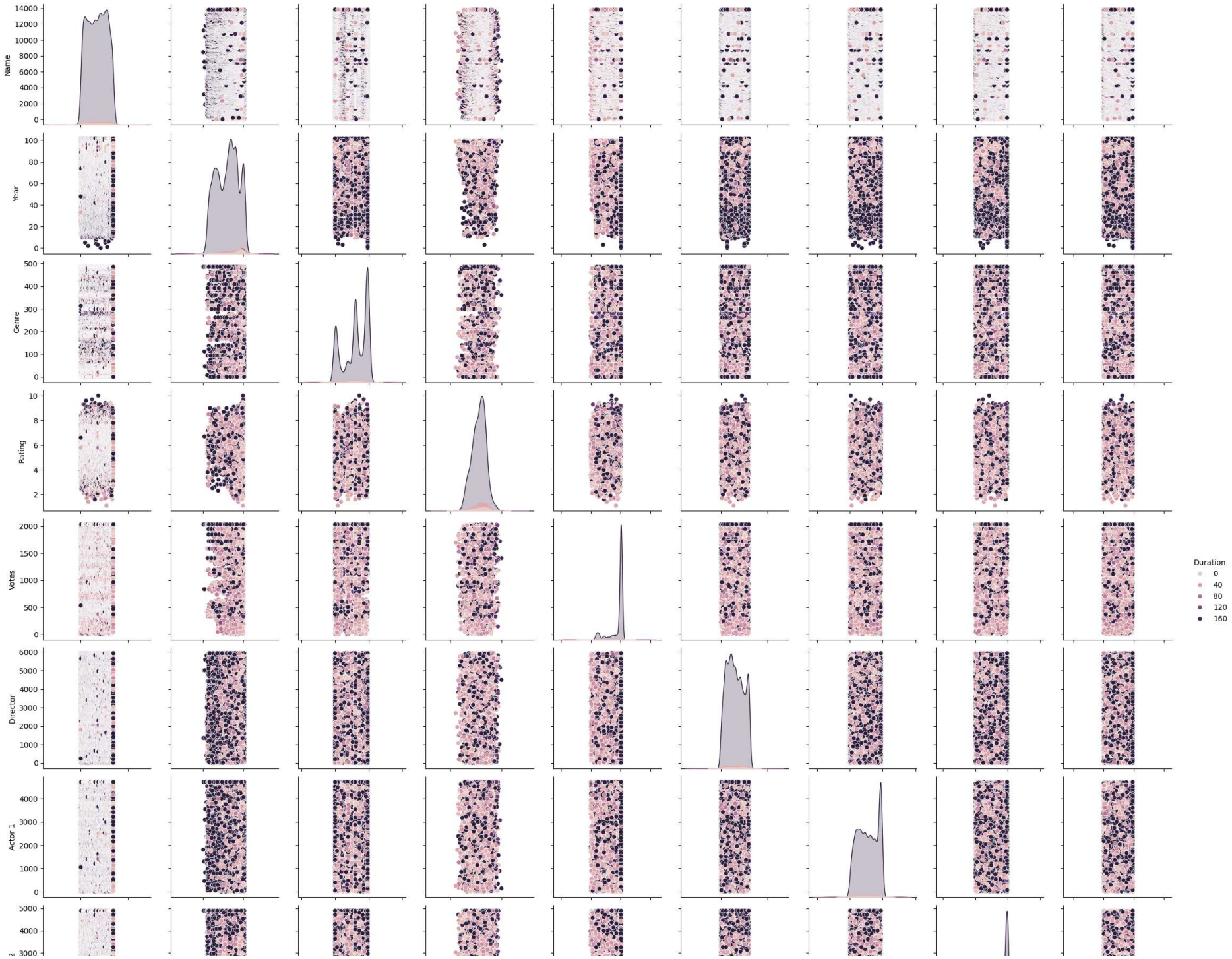
```
ohe.fit_transform(dat[["Name"]]).toarray()
```

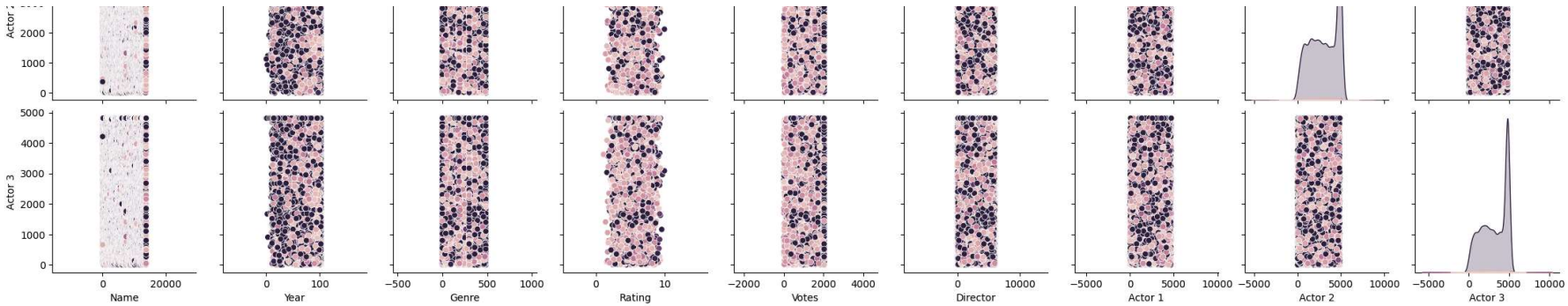
```
array([[1., 0., 0., ..., 0., 0., 0.],
       [0., 1., 0., ..., 0., 0., 0.],
       [0., 0., 1., ..., 0., 0., 0.],
       ...,
```

```
       [0., 0., 0., ..., 1., 0., 0.],
       [0., 0., 0., ..., 0., 1., 0.],
       [0., 0., 0., ..., 0., 0., 0.]])
```

```
import seaborn as sns
sns.pairplot(dat,hue="Duration")
```

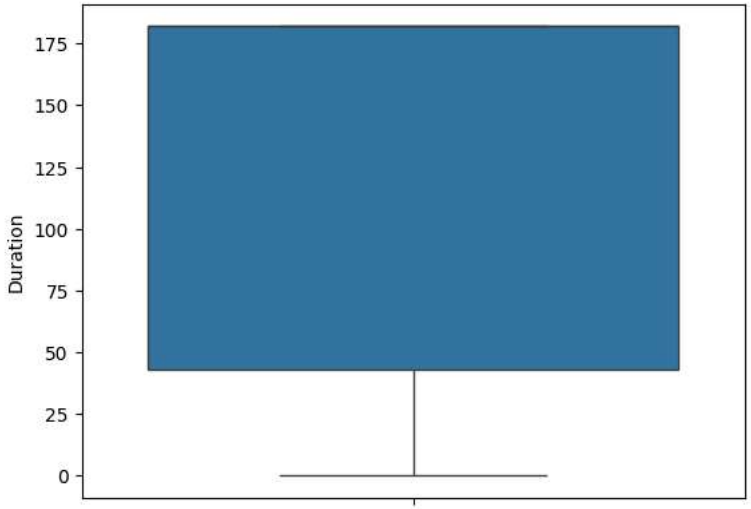
```
<seaborn.axisgrid.PairGrid at 0x7c63f59b0950>
```




Actor 3
Name
Year
Genre
Rating
Votes
Director
Actor 1
Actor 2
Actor 3

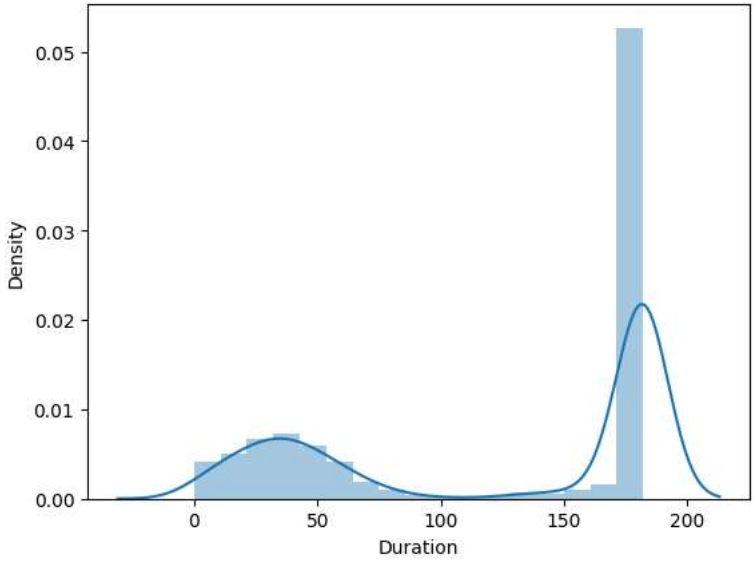
```
sns.boxplot(dat["Duration"])
```

```
<Axes: ylabel='Duration'>
```



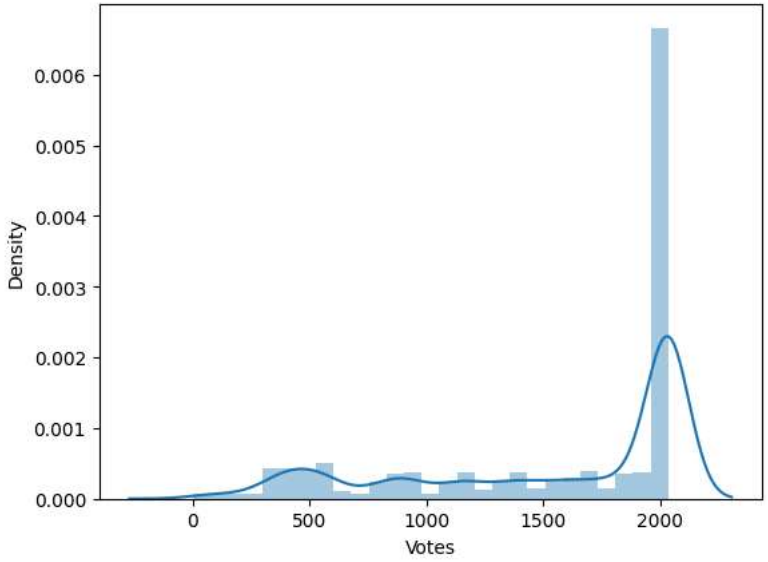

```
sns.distplot(dat["Duration"])
```

```
<ipython-input-21-dddfa5794f79>:1: UserWarning:

`distplot` is a deprecated function and will be removed in seaborn v0.14.0.

Please adapt your code to use either `displot` (a figure-level function with
similar flexibility) or `histplot` (an axes-level function for histograms).

For a guide to updating your code to use the new functions, please see
https://gist.github.com/mwaskom/de44147ed2974457ad6372750bbe5751

  sns.distplot(dat["Duration"])
<Axes: xlabel='Duration', ylabel='Density'>
```




```
sns.distplot(dat["Votes"])
```

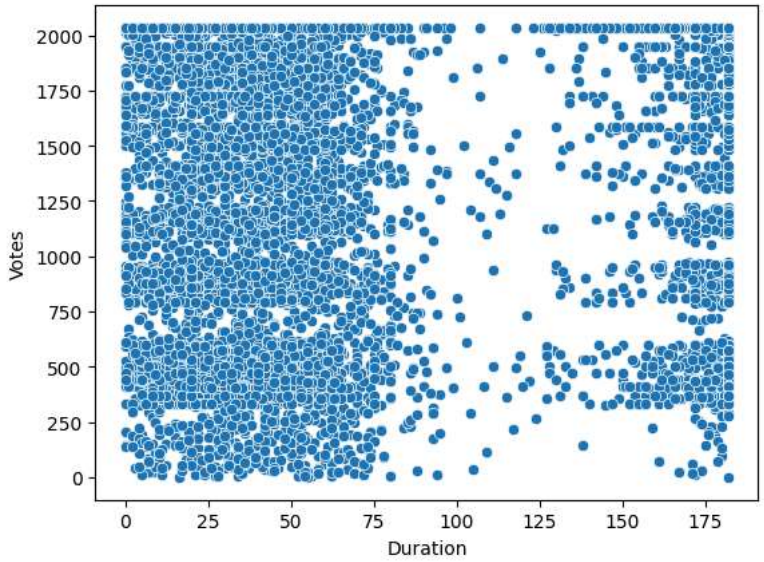
```
<ipython-input-22-07434a9fb940>:1: UserWarning:

`distplot` is a deprecated function and will be removed in seaborn v0.14.0.

Please adapt your code to use either `displot` (a figure-level function with
similar flexibility) or `histplot` (an axes-level function for histograms).

For a guide to updating your code to use the new functions, please see
https://gist.github.com/mwaskom/de44147ed2974457ad6372750bbe5751

  sns.distplot(dat["Votes"])
<Axes: xlabel='Votes', ylabel='Density'>
```



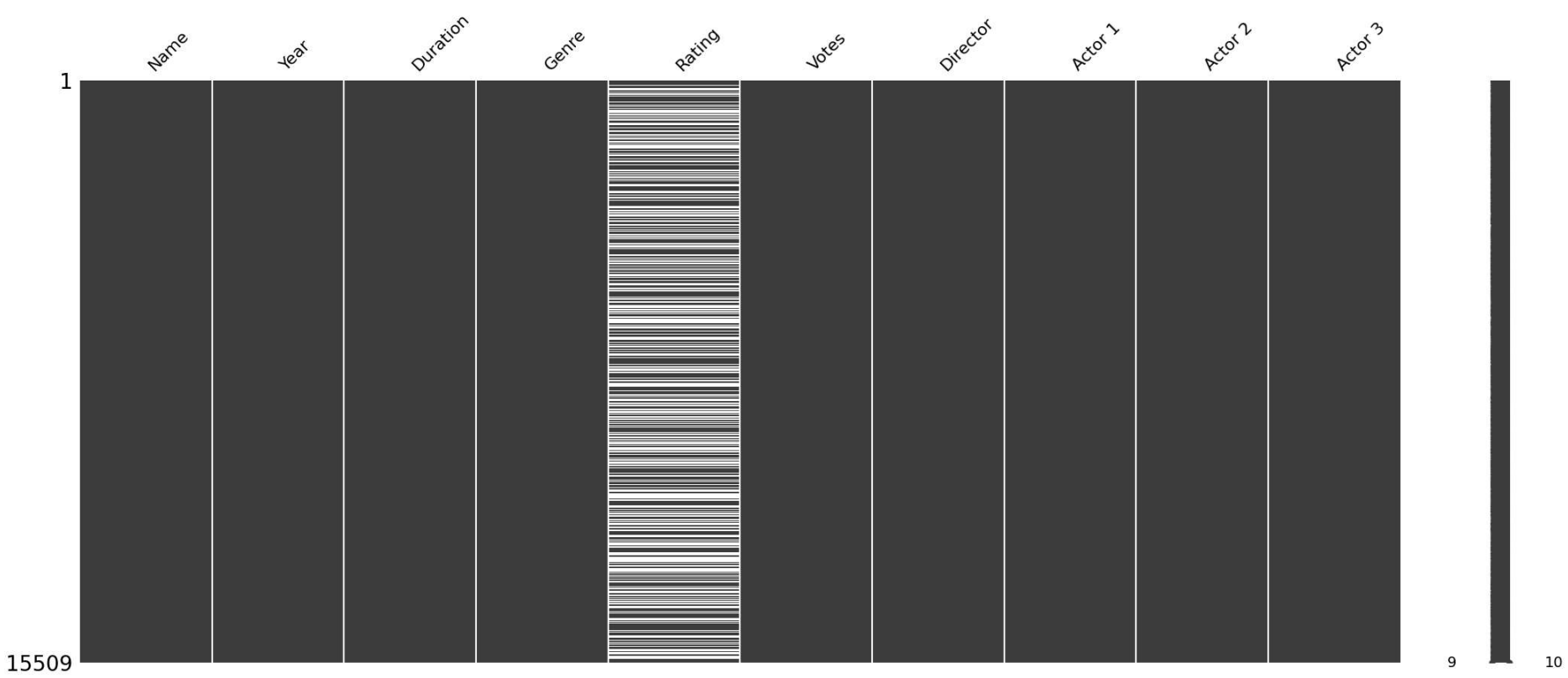
Double-click (or enter) to edit

```
sns.scatterplot(x="Duration",y="Votes",data=dat)
```

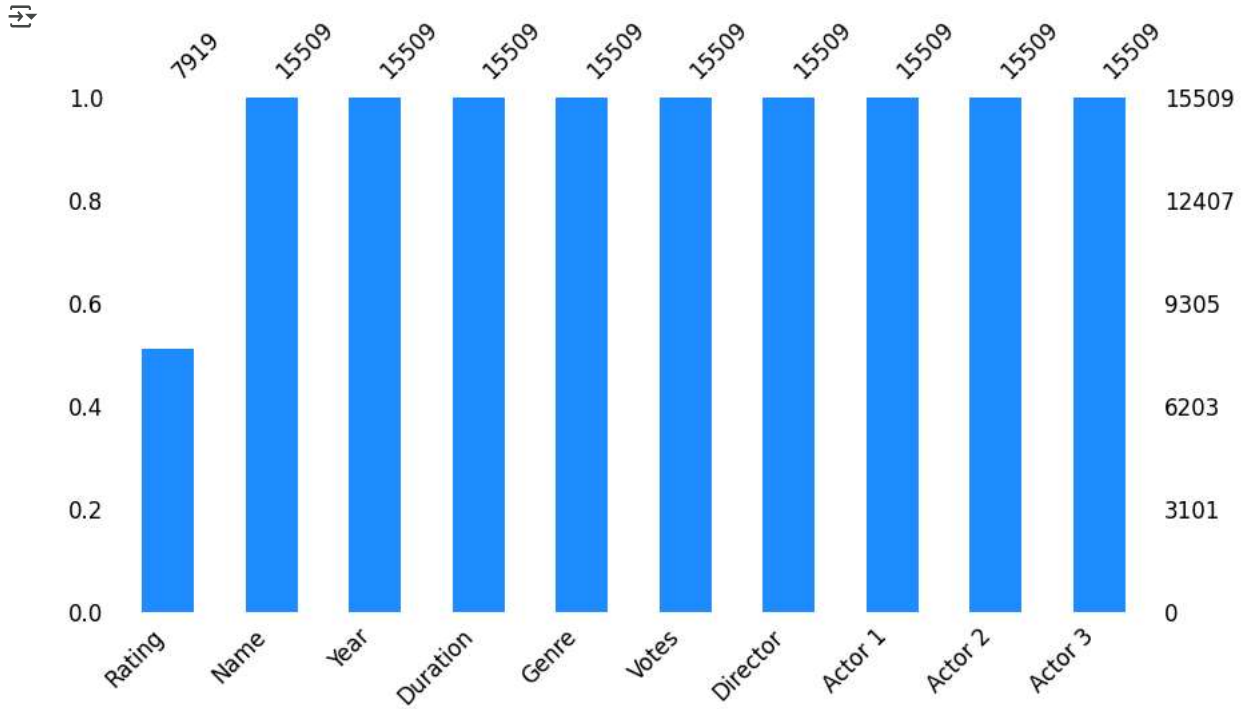
```
<Axes: xlabel='Duration', ylabel='Votes'>
```



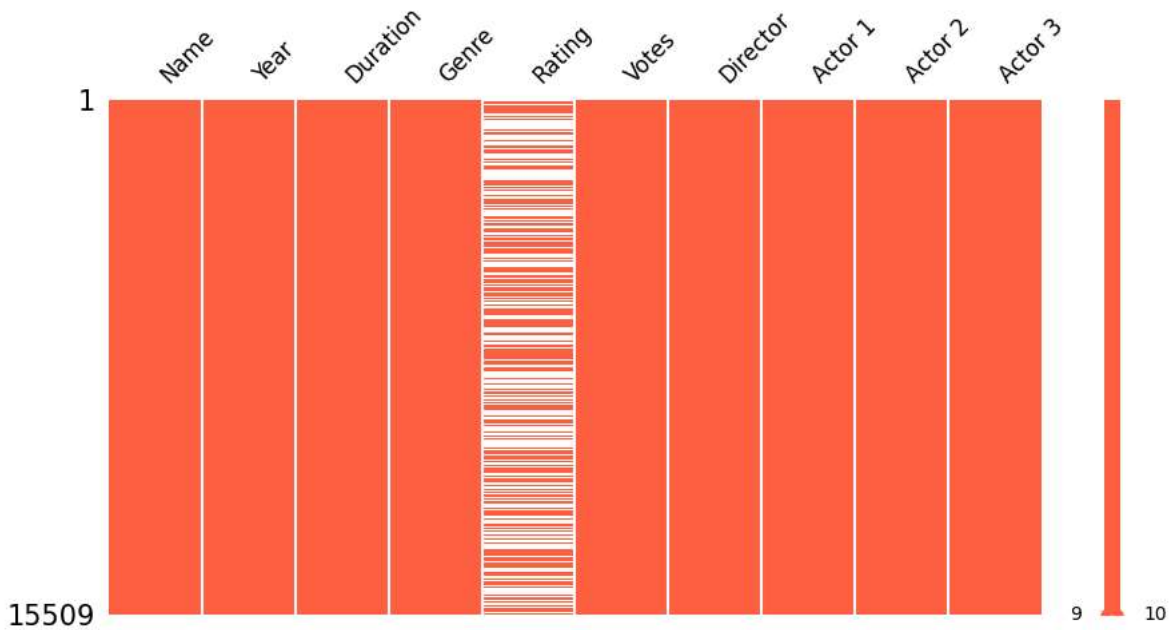

```
dat['Duration'] = dat['Duration'].astype(int)
import missingno as msno
msno.matrix(dat)
plt.show()
```




```
import missingno
missingno.bar(dat, color="dodgerblue", sort="ascending", figsize=(10,5), fontsize=12);
```

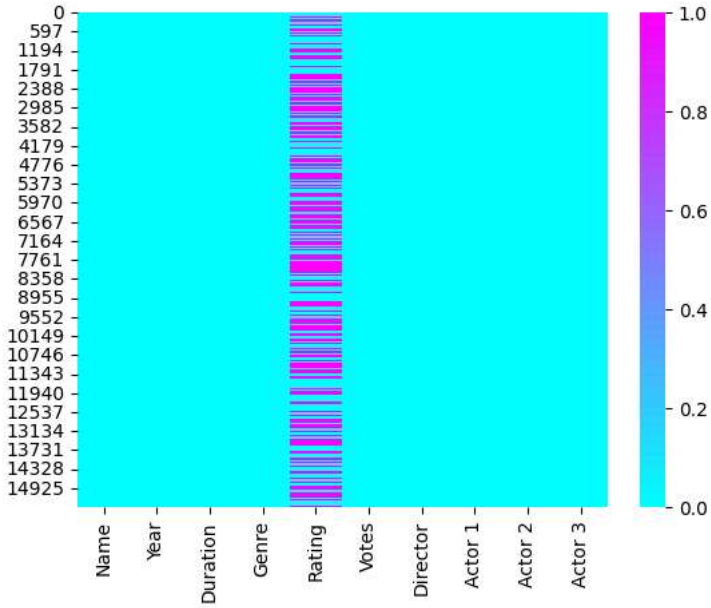



```
missingno.matrix(dat, figsize=(10,5), fontsize=12, color=(1, 0.38, 0.27));
```




```
missingno.heatmap(dat, figsize=(10,5), fontsize=12);
```


```
sns.heatmap(dat.isnull(),cmap='cool');
```




```
X = dat.iloc[ :, :-1]
Y = dat.iloc[ :, -1]
```

```
Y.head(2)
```

	Actor 3
0	3108
1	527

dtype: int64

```
from sklearn.model_selection import train_test_split
X_train, X_test, Y_train, Y_test = train_test_split(X, Y, test_size = 0.2, random_state = 42)
```