Movies Dataset from Pirated Sites

March 27, 2024

[1]:
```
import pandas as pd
import matplotlib.pyplot as plt
```

0.0.1 movies_dataset.csv

[2]:
```
data = pd.read_csv('data/movies_dataset.csv')
```

[3]:
```
type(data)
```

[3]:
```
pandas.core.frame.DataFrame
```

[4]:
```
data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 20548 entries, 0 to 20547
Data columns (total 15 columns):
 #   Column           Non-Null Count  Dtype  
---  ------           --------------  -----  
 0   Unnamed: 0       20548 non-null  int64  
 1   IMDb-rating      19707 non-null  float64
 2   appropriate_for  11072 non-null  object 
 3   director         18610 non-null  object 
 4   downloads        20547 non-null  object 
 5   id               20548 non-null  int64  
 6   industry         20547 non-null  object 
 7   language         20006 non-null  object 
 8   posted_date      20547 non-null  object 
 9   release_date     20547 non-null  object 
 10  run_time         18780 non-null  object 
 11  storyline        18847 non-null  object 
 12  title            20547 non-null  object 
 13  views            20547 non-null  object 
 14  writer           18356 non-null  object 
dtypes: float64(1), int64(2), object(12)
memory usage: 2.4+ MB
```

0.0.2

[5]:
```
s = data['appropriate_for']
s.value_counts()
```

[5]:
```
R                4384
Not Rated        2142
PG-13            1968
PG                886
TV-14             694
TV-MA             406
G                 152
Unrated           132
TV-PG             115
TV-G               99
TV-Y7              45
TV-Y               25
Approved            9
NC-17               4
TV-Y7-FV            3
Passed              3
MA-17               1
TV-13               1
Drama               1
Drama, Romance      1
18+                 1
Name: appropriate_for, dtype: int64
```

[6]:
```
s = data['director']
s.value_counts()
```

[6]:
```
Venky Atluri                                   405
Simone Stock                                   403
Xavier Manrique                                403
John Swab                                      205
Neil Jordan                                    205
                                              ...
Agnieszka Smoczynska                             1
Dylan Thomas Ellis                               1
Sunil Thakur, Sunil Dhawan, Shivani Thakur       1
Suman Mukhopadhyay                               1
Shea Sizemore                                    1
Name: director, Length: 9672, dtype: int64
```

[7]:
```
s = data['industry']
s.value_counts()
```

[7]:
```
Hollywood / English    14649
Bollywood / Indian      2645
Tollywood               1172
Anime / Kids            1049
Wrestling                433
Punjabi                  332
Stage shows              129
Pakistani                 92
Dub / Dual Audio          45
3D Movies                  1
Name: industry, dtype: int64
```

[8]:
```
s = data['language']
s.value_counts()
```

[8]:
```
English                                  12657
Hindi                                     2558
English,Spanish                            391
Punjabi                                    310
English,Hindi                              304
                                          ...
English,Korean,Spanish                       1
Norwegian,Swedish                            1
Spanish,Chinese,English,Maori,French         1
Urdu,Punjabi,English                         1
Spanish,German,English                       1
Name: language, Length: 1168, dtype: int64
```

[9]:
```
s = data['posted_date']
s.value_counts()
```

[9]:
```
13 Feb, 2023    812
20 Feb, 2023    607
15 Feb, 2023    607
10 Feb, 2023    485
16 Feb, 2023    406
                ...
12 Sep, 2009      1
08 Sep, 2009      1
01 Sep, 2009      1
18 Aug, 2009      1
30 Nov, 2011      1
Name: posted_date, Length: 4123, dtype: int64
```

[10]:
```
s = data['release_date']
s.value_counts()
```

[10]:
```
Jan 01 1970    962
Feb 03 2023    616
Feb 17 2023    607
Feb 10 2023    410
Feb 11 2023    402
               …
Sep 05 2003      1
Dec 29 2022      1
Aug 24 2013      1
Jan 12 2014      1
Mar 28 1958      1
Name: release_date, Length: 4886, dtype: int64
```

[11]:
```
s = data['storyline']
s.value_counts()
```

[11]:
```
The life of a young man and his struggles against the privatization of
education.
402
Follows\r\n a New York City family hiding out in the Hamptons whose bubble is
\r\npopped when a Bloody Mary-swilling, pot-smoking 'Charlie' comes to bring\r\n
a lifetime of hurt that might heal them all.
402
It follows Kara Robinson as she survives an abduction and ultimately brings down
a serial killer.
402
Doc\r\n facilitates a fragile truce between the Governor and Cartel, trading
\r\nprosecutorial leniency for finance. With no more truce, Doc is left to
\r\nfend for himself and protect the one untainted thing in his life: his
\r\ndaughter, Little Dixie.
202
A\r\n young, gay Black man, rejected by his mother and with few options for
\r\nhis future, decides to join the Marines, doing whatever it takes to
\r\nsucceed in a system that would cast him aside.
202
                                                                          …
Four waves of increasingly deadly attacks have left most of Earth in ruin.
Against a backdrop of fear and distrust, Cassie is on the run, desperately
trying to save her younger brother. As she prepares for the inevitable and
lethal fifth wave, Cassie teams up with a young man who may become her final
hope - if she can only trust him.
1
Yamuna along with her son Laxman locates to Mumbai leaving behind her abusive
husband.She takes shelter in the house of her aunt Chandra whom she calls
Akka.Yamuna's only aim is to give a better education to her son.Chandra finds
her a job as sweeper in a art school.Yamuna finds that Chandra poses as a nude
model to the students of the school.Chandra confines Yamuna to take up the job
```

```
being nude out there the students don't look at you in lust but as a project.
1
A young violinist struggles to assert her individuality amidst the intense
pressure of her pianist father, and the weight of her own musical ability.
1
A right wing talk show host's life takes a sudden turn when his 16 year old
niece comes crashing into his life.
1
While driving his car on a rainy night, Anand's car breaks down, and he goes to
seek shelter in a nearby house. He is let into the house by the servant, and he
is permitted to stay until the rains stop be able to get his car fixed. It is
here that he will find out about his previous birth, his true love, Madhumati,
their ill-fated, star-crossed and tragic romance, and how events in his previous
birth are going to effect him in this life-time.                          1
Name: storyline, Length: 15748, dtype: int64
```

[12]:
```
s = data['title']
s.value_counts()
```

[12]:
```
The Girl Who Escaped: The Kara Robinson Story    402
Vaathi                                           402
Who Invited Charlie?                             402
Little Dixie                                     202
The Inspection                                   202
                                                ...
Kesari                                             1
Old Boys                                           1
American Exit                                      1
Adventures of Aladdin                              1
Madhumati                                          1
Name: title, Length: 16572, dtype: int64
```

[13]:
```
s = data['writer']
s.value_counts()
```

[13]:
```
Nicholas Schutt                          403
Venky Atluri                             402
Haley Harris                             402
John Swab                                205
Elegance Bratton                         202
                                        ...
Barbara Samuels, Joseph Boyden             1
Maria Allred                               1
Pia Mechler                                1
Paul Flannery, David Ryan Keith            1
Khwaja Ahmad Abbas, Khwaja Ahmad Abbas     1
Name: writer, Length: 13603, dtype: int64
```

0.0.3 5

5

[14]:
```
data.describe()
```

[14]:

	Unnamed: 0	IMDb-rating	id
count	20548.000000	19707.000000	20548.000000
mean	10273.500000	5.762151	222351.199776
std	5931.841001	1.374041	138422.327931
min	0.000000	1.100000	1.000000
25%	5136.750000	4.800000	96122.250000
50%	10273.500000	5.700000	264457.500000
75%	15410.250000	6.600000	354561.250000
max	20547.000000	9.900000	372092.000000

[15]:
```
data['IMDb-rating'].isnull().sum()
```

[15]: 841

[16]:
```
data['downloads'].isnull().sum()
```

[16]: 1

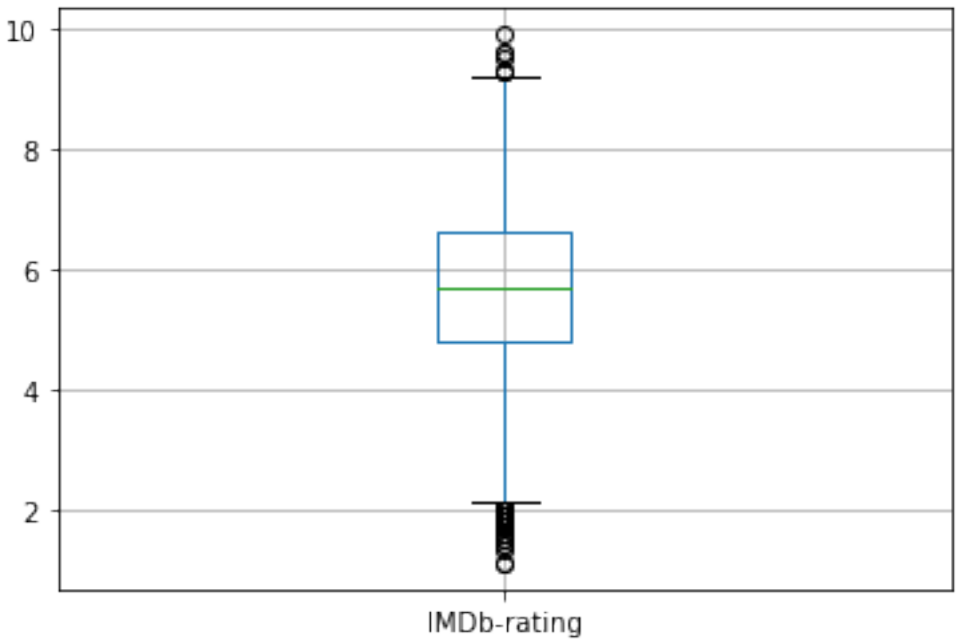
[17]:
```
data['views'].isnull().sum()
```

[17]: 1

0.0.4

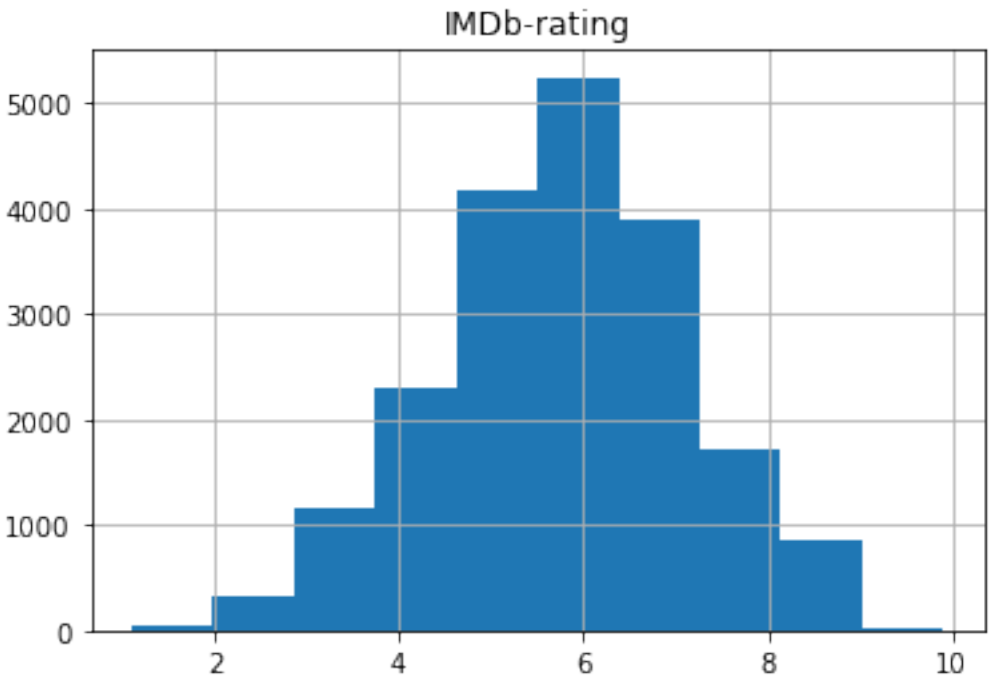
[18]:
```
data.boxplot(column='IMDb-rating')
```

[18]: <AxesSubplot:>



[19]:
```
data.hist(column='IMDb-rating')
```

[19]:
```
array([[<AxesSubplot:title={'center':'IMDb-rating'}>]], dtype=object)
```



0.0.5

[20]:
```
data['appropriate_for'].isnull().sum()
```

[20]:
```
9476
```

[21]:
```
data.shape
```

[21]:
```
(20548, 15)
```

[22]:
```
del data['appropriate_for']
```

[23]:
```
data.shape
```

[23]:
```
(20548, 14)
```

[24]:
```
data['director']
```

[24]:
```
0           John Swab
1         Paul Ziller
2        Ben Wheatley
3        Venky Atluri
4        Shaji Kailas
             …
20543             NaN
20544       Bimal Roy
20545             NaN
20546             NaN
20547             NaN
Name: director, Length: 20548, dtype: object
```

[25]:
```
#
mode = data['director'].value_counts().index[0]
```

[26]:
```
#
data['director'].fillna(mode, inplace=True)
```

[27]:
```
data['director']
```

[27]:
```
0           John Swab
1         Paul Ziller
2        Ben Wheatley
3        Venky Atluri
4        Shaji Kailas
             …
20543    Venky Atluri
20544       Bimal Roy
20545    Venky Atluri
```

```
20546    Venky Atluri
20547    Venky Atluri
Name: director, Length: 20548, dtype: object
```

[28]:
```
data['IMDb-rating']
```

[28]:
```
0        4.8
1        6.4
2        5.2
3        8.1
4        4.6
        …
20543    NaN
20544    7.7
20545    8.0
20546    NaN
20547    NaN
Name: IMDb-rating, Length: 20548, dtype: float64
```

[29]:
```
#
data['IMDb-rating'].fillna(data['IMDb-rating'].mean(), inplace=True)
```

[30]:
```
data['IMDb-rating']
```

[30]:
```
0        4.800000
1        6.400000
2        5.200000
3        8.100000
4        4.600000
           …
20543    5.762151
20544    7.700000
20545    8.000000
20546    5.762151
20547    5.762151
Name: IMDb-rating, Length: 20548, dtype: float64
```

[31]:
```
data['run_time'].isnull().sum()
```

[31]:
```
1768
```

[32]:
```
data['run_time']
```

[32]:
```
0          105
1           84
2        1h 47min
3          139
```

```
4             122
            …
20543         NaN
20544         159
20545    1h 50min
20546         NaN
20547         NaN
Name: run_time, Length: 20548, dtype: object
```

[33]:
```
#
data['run_time'].fillna(method='pad', inplace=True)
```

[34]:
```
data['run_time']
```

[34]:
```
0             105
1              84
2        1h 47min
3             139
4             122
            …
20543    2h 36min
20544         159
20545    1h 50min
20546    1h 50min
20547    1h 50min
Name: run_time, Length: 20548, dtype: object
```

[ ]:
```

```