Task 1: Movie Rating Prediction

Going to build a model that predicts the rating of a movie based on features like actors, directors, and genres

In [4]:
```
import numpy as bp
import os
import pandas as pd
import matplotlib.pyplot as plt
%matplotlib inline
from matplotlib import style
import seaborn as sns
import heapq
```

In [16]:
```
#import csv file
df = pd.read_csv('Movies.csv')
```

Data Cleaning Process

In [74]:
```
#Check the overall data details
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Index: 5659 entries, 1 to 15508
Data columns (total 10 columns):
 #   Column    Non-Null Count  Dtype
---  ------    --------------  -----
 0   Name      5659 non-null   object
 1   Year      5659 non-null   int32
 2   Duration  5659 non-null   object
 3   Genre     5659 non-null   object
 4   Rating    5659 non-null   int32
 5   Votes     5659 non-null   object
 6   Director  5659 non-null   object
 7   Actor 1   5659 non-null   object
 8   Actor 2   5659 non-null   object
 9   Actor 3   5659 non-null   object
dtypes: int32(2), object(8)
memory usage: 442.1+ KB
```

In [19]:
```
#check the null values
pd.isnull(df).sum()
```

Out[19]:
```
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
```

In [20]:
```
df.shape
```

Out[20]:
```
(15509, 10)
```

In [21]:
```
#Remove the null values
df.dropna(inplace=True)
```

In [22]:
```
df.shape
```

Out[22]:
```
(5659, 10)
```

In [23]:
```
#Null values removed
pd.isnull(df).sum()
```

Out[23]:
```
Name        0
Year        0
Duration    0
Genre       0
Rating      0
Votes       0
Director    0
Actor 1     0
Actor 2     0
Actor 3     0
dtype: int64
```

In [25]:
```
#change from float to int
df['Rating'] = df['Rating'].astype('int')
```

In [145…
```
df['Duration'] = df['Duration'].astype('int')
```

In [102…
```
df['Year'] = df['Year'].astype(str).str.lstrip('<').str.rstrip('+')
```

In [28]:
```
df.columns
```

Out[28]:
```
Index(['Name', 'Year', 'Duration', 'Genre', 'Rating', 'Votes', 'Director',
       'Actor 1', 'Actor 2', 'Actor 3'],
      dtype='object')
```

In [106…
```
#Remove the special character in front of the years
df['Year'] = [x.strip("-") for x in df['Year']]
```

In [129…
```
df['Votes'] = df['Votes'].str.replace(',','').astype(int)
```

In [130…
```
df['Name'] = df['Name'].str.replace(r"[\"\'\|\=\.\...\:\<\>\@\#\*\?\$\%\^\&\!\_\-\;\'\-\,]", '')
```

In [53]:
```
df['Name'] = [x.strip(":",) for x in df['Name']]
```

In [48]:
```
df['Name'] = [x.strip("#",) for x in df['Name']]
```

In [55]:
```
df['Name'] = [x.strip("?",) for x in df['Name']]
```

In [56]:
```
df['Name'] = [x.strip(":",) for x in df['Name']]
```

In [61]:
```
df['Name'] = [x.strip("@",) for x in df['Name']]
```

In [40]:
```
df['Name'] = [x.strip("...") for x in df['Name']]
```

In [38]:
```
df['Duration'] = [x.strip("min",) for x in df['Duration']]
```

In [108…
```
#Drop duplicate values
df.drop_duplicates(subset = 'Name', keep = "first").head(2)
```

Out[108…

	Name	Year	Duration	Genre	Rating	Votes	Director	Actor 1	Actor 2	Actor 3
1	Gadhvi (He thought he was Gandhi)	2019	109	Drama	7	8	Gaurav Bakshi	Rasika Dugal	Vivek Ghamande	Arvind Jangid
3	Yaaram	2019	110	Comedy, Romance	4	35	Ovais Khan	Prateik	Ishita Raj	Siddhant Kapoor

In [107…
```
#change the name of column
df.rename(columns={'Duration':'Duration (min)'}).head(2)
```

Out[107…

	Name	Year	Duration (min)	Genre	Rating	Votes	Director	Actor 1	Actor 2	Actor 3
1	Gadhvi (He thought he was Gandhi)	2019	109	Drama	7	8	Gaurav Bakshi	Rasika Dugal	Vivek Ghamande	Arvind Jangid
3	Yaaram	2019	110	Comedy, Romance	4	35	Ovais Khan	Prateik	Ishita Raj	Siddhant Kapoor

In [109…
```
#Rename the column of duration
df.rename(columns={'Duration':'Duration (min)'}).head(2)
```

Out[109…

	Name	Year	Duration (min)	Genre	Rating	Votes	Director	Actor 1	Actor 2	Actor 3
1	Gadhvi (He thought he was Gandhi)	2019	109	Drama	7	8	Gaurav Bakshi	Rasika Dugal	Vivek Ghamande	Arvind Jangid
3	Yaaram	2019	110	Comedy, Romance	4	35	Ovais Khan	Prateik	Ishita Raj	Siddhant Kapoor

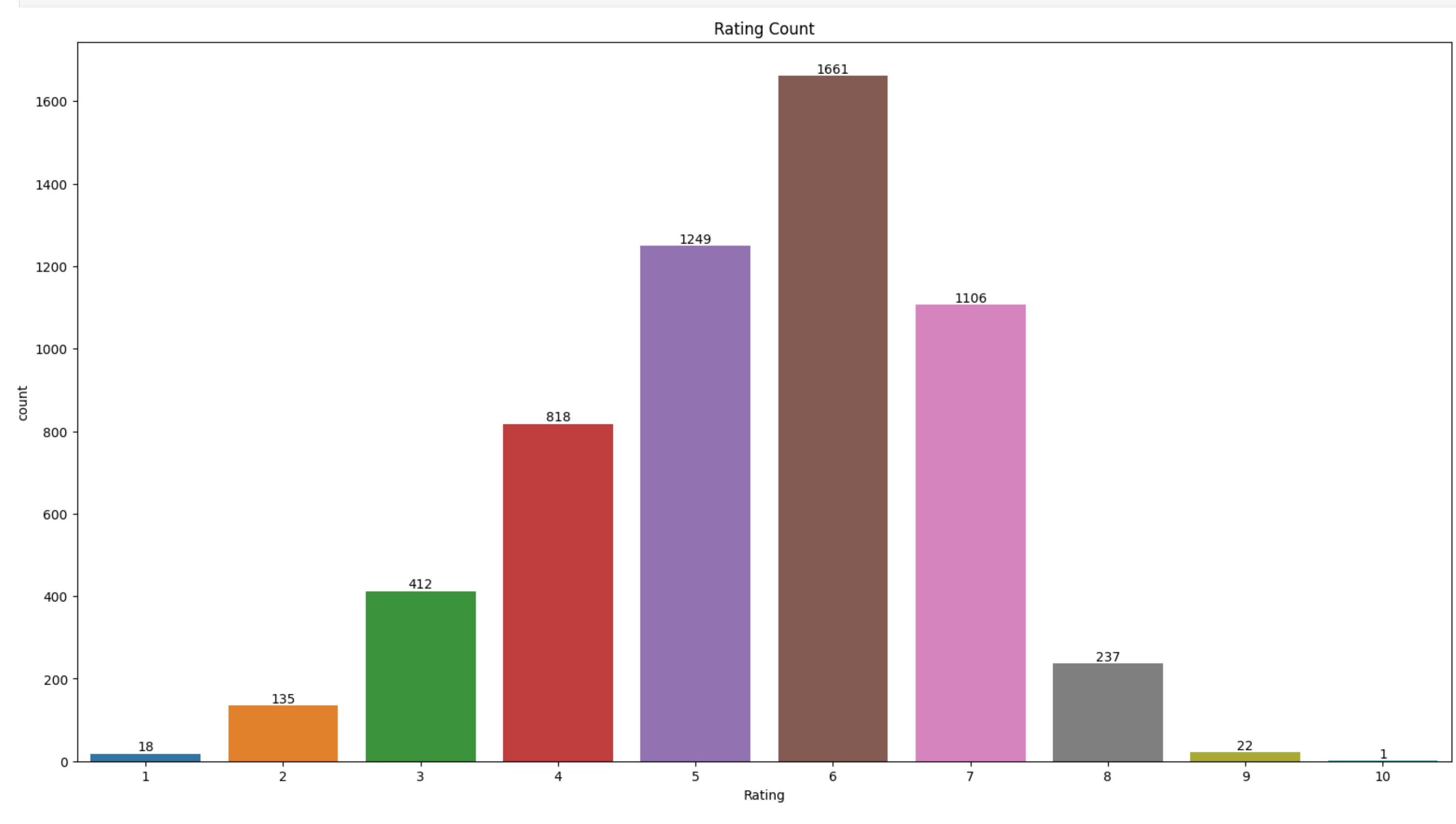
Exploratory Data Analysis

In [110…
```
import warnings
warnings.filterwarnings('ignore')
```

Rating Counts

In [111…
```
plt.figure(figsize = (19,10))
ax = sns.countplot(x = 'Rating', data = df)

for bars in ax. containers:
    ax.bar_label(bars)
plt.title('Rating Count')
plt.show()
```



In [136…
```
df['Rating'].max()
```

Out[136…
```
10
```

In [135…
```
df['Rating'].min()
```

Out[135…
```
1
```

Maximum Duration of movie

In [149…
```
df['Duration'].max()
```

Out[149…
```
321
```

In [280…
```
df['Duration'].max()
df['Name'][df['Duration'].idxmax()]
```

Out[280…
```
'Gangs of Wasseypur'
```

Minimum Voting

In [133…
```
df['Votes'].min()
```

Out[133…
```
5
```

Highest Vote to which one movie?

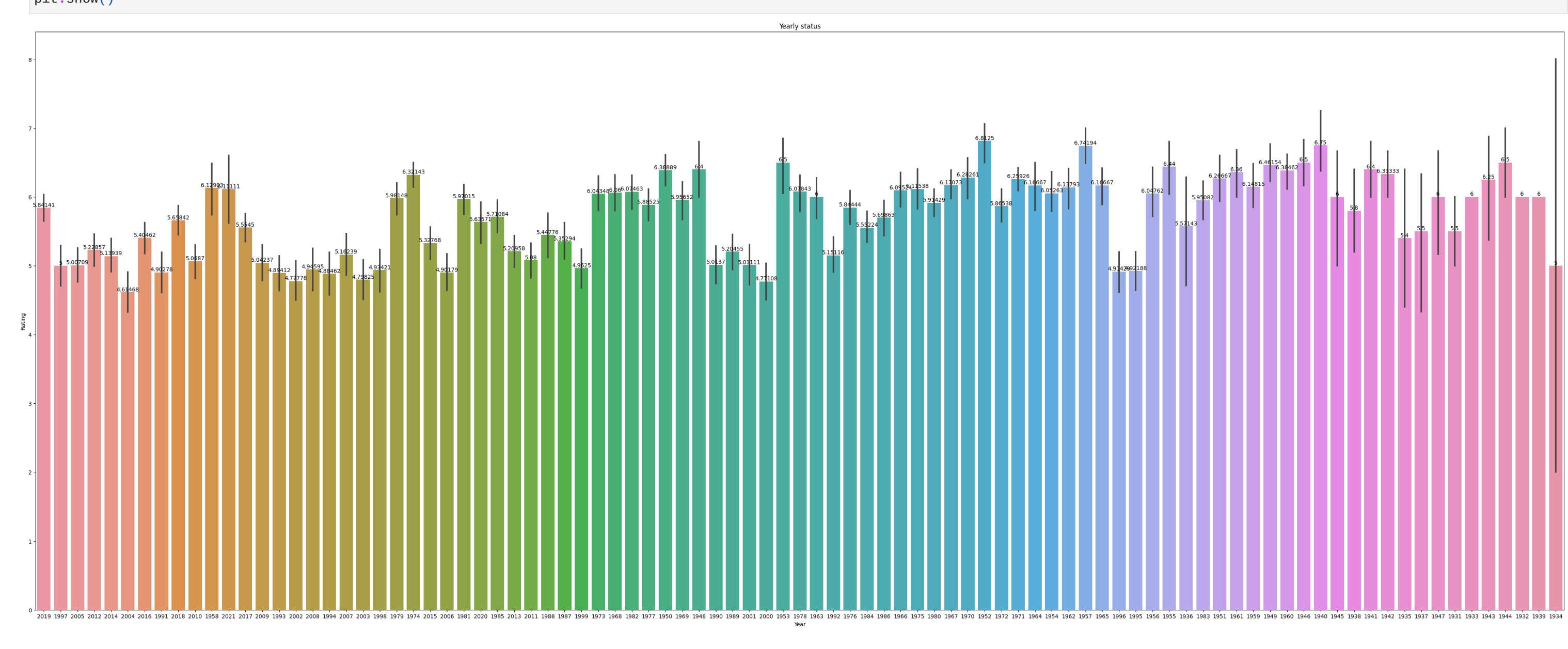
In [142…
```
print(df['Votes'].max())
df['Name'][df['Votes'].idxmax()]
```

```
591417
```

Out[142…
```
'Life of Pi'
```

Which year has highest Rating?

In [138…
```
plt.figure(figsize = (50,19))
ax = sns.barplot(x = 'Year', y= 'Rating' ,data = df)
for bars in ax. containers:
    ax.bar_label(bars)
plt.title('Yearly status')
plt.show()
```



Top 10 movies with highest voting

In [210…
```
df.nlargest(10, columns=['Votes'], keep='first')
```

Out[210…

	Name	Year	Duration	Genre	Rating	Votes	Director	Actor 1	Actor 2	Actor 3
8219	Life of Pi	2012	127	Adventure, Drama, Fantasy	7	591417	Ang Lee	Suraj Sharma	Irrfan Khan	Adil Hussain
75	3 Idiots	2009	170	Comedy, Drama	8	357889	Rajkumar Hirani	Aamir Khan	Madhavan	Mona Singh
8233	Lion	2016	118	Biography, Drama	8	220526	Garth Davis	Dev Patel	Nicole Kidman	Rooney Mara
4848	Gandhi	1982	191	Biography, Drama, History	8	220118	Richard Attenborough	Ben Kingsley	John Gielgud	Rohini Hattangadi
14038	The Darjeeling Limited	2007	91	Adventure, Comedy, Drama	7	185127	Wes Anderson	Owen Wilson	Adrien Brody	Jason Schwartzman
8228	Like Stars on Earth	2007	165	Drama, Family	8	175810	Aamir Khan	Amole Gupte	Darsheel Safary	Aamir Khan
10882	PK	2014	153	Comedy, Drama, Musical	8	168150	Rajkumar Hirani	Aamir Khan	Anushka Sharma	Sanjay Dutt
3410	Dangal	2016	161	Action, Biography, Drama	8	165074	Nitesh Tiwari	Aamir Khan	Sakshi Tanwar	Fatima Sana Shaikh
11463	Radhe	2021	135	Action, Crime, Thriller	1	162455	Prabhu Deva	Salman Khan	Disha Patani	Randeep Hooda
3829	Dil Bechara	2020	101	Comedy, Drama, Romance	6	117377	Mukesh Chhabra	Sushant Singh Rajput	Sanjana Sanghi	Sahil Vaid

Movie's duration having less than 100 minutes.

In [201…
```
df[(df.Duration <= 100)]
```

Out[201…

	Name	Year	Duration	Genre	Rating	Votes	Director	Actor 1	Actor 2	Actor 3
8	A Question Mark	2012	82	Horror, Mystery, Thriller	5	326	Allyson Patel	Yash Dave	Muntazir Ahmad	Kiran Bhatia
10	1:1.6 An Ode to Lost Love	2004	96	Drama	6	17	Madhu Ambat	Rati Agnihotri	Gulshan Grover	Atul Kulkarni
18	10ml LOVE	2010	87	Comedy, Drama, Romance	6	162	Sharat Katariya	Neil Bhoopalam	Anusha Bose	Manu Rishi Chadha
36	19 Revolutions	2004	94	Drama	4	16	Sridhar Reddy	Tarun Arora	Gulshan Grover	Vinay Pandey
50	2 Nights in Soul Valley	2012	80	Adventure, Horror, Mystery	6	21	Harish Sharma	Hemant Pandey	Sumeet Sharma	Sumeet Sharma
...	...	...	...	...	...	...	...	...	...	...
15229	Yeh Hai Chakkad Bakkad Bumbe Bo	2003	90	Adventure	6	9	Sridhar Rangayan	Tom Alter	Mona Ambegaonkar	Aardra Athalye
15288	Yeh Suhaagraat Impossible	2019	92	Comedy	8	500	Abhinav Thakur	Preetika Chauhan	Aarav Mavi	Aloknath Pathak
15312	Yours Truly	2018	84	Drama	5	102	Sanjoy Nag	Soni Razdan	Aahana Kumra	Pankaj Tripathi
15320	Yug the law of karma	2021	75	Action, Crime, Drama	7	10	Dilip Kumar	Vinod Kumar	Dilip Kumar	Saurav Bagga
15488	Zoo	2018	100	Drama	5	78	Shlok Sharma	Shashank Arora	Prince Daniel	Shatakshi Gupta

596 rows × 10 columns

Movie's duration having less than 50 minutes

In [279…
```
df['Duration'].loc[lambda x:x < 50]
```

Out[279…
```
280      46
1513     48
2723     49
4504     45
5756     47
8046     48
8709     45
10504    49
10974    21
12002    45
12658    47
13217    45
14071    48
Name: Duration, dtype: int32
```

Top 5 movies with highest-rated

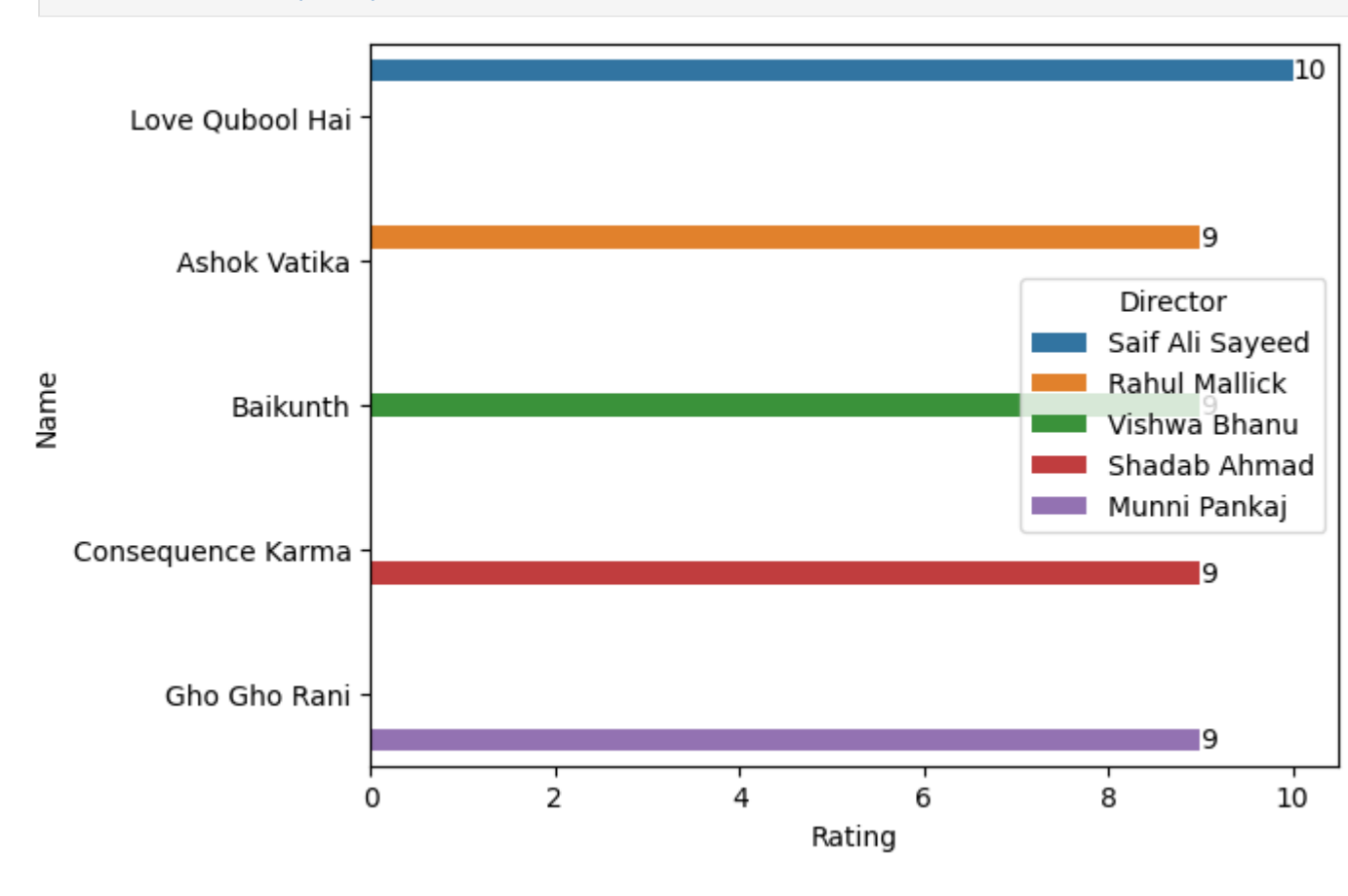
In [287…
```
Top5_movies=df.nlargest(5,'Rating')[['Name', 'Votes', 'Rating', 'Director']]\
.set_index('Name')
```

In [288…
```
Top5_movies
```

Out[288…

Name	Votes	Rating	Director
Love Qubool Hai	5	10	Saif Ali Sayeed
Ashok Vatika	7	9	Rahul Mallick
Baikunth	29	9	Vishwa Bhanu
Consequence Karma	223	9	Shadab Ahmad
Gho Gho Rani	47	9	Munni Pankaj

In [234…
```
ax = sns.barplot(x='Rating', y=Top5_movies.index,data=Top5_movies, hue='Director')
for bars in ax. containers:
    ax.bar_label(bars)
```



Top 5 movies with highest-voted

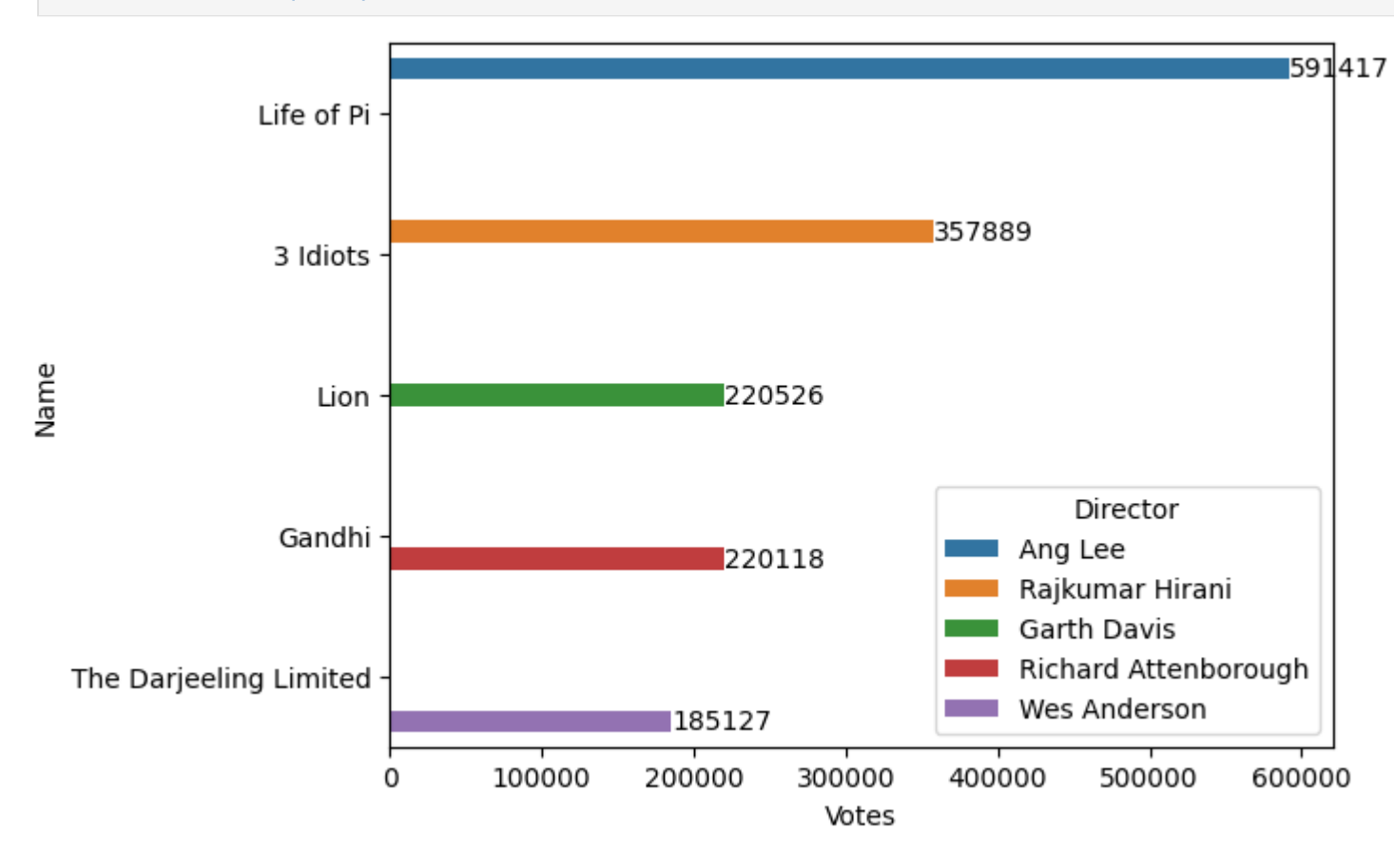
In [285…
```
Top5_movies=df.nlargest(5,'Votes')[['Name', 'Year', 'Rating','Votes', 'Director','Actor 1','Actor 2', 'Actor 3']]\
.set_index('Name')
```

In [286…
```
Top5_movies
```

Out[286…

Name	Year	Rating	Votes	Director	Actor 1	Actor 2	Actor 3
Life of Pi	2012	7	591417	Ang Lee	Suraj Sharma	Irrfan Khan	Adil Hussain
3 Idiots	2009	8	357889	Rajkumar Hirani	Aamir Khan	Madhavan	Mona Singh
Lion	2016	8	220526	Garth Davis	Dev Patel	Nicole Kidman	Rooney Mara
Gandhi	1982	8	220118	Richard Attenborough	Ben Kingsley	John Gielgud	Rohini Hattangadi
The Darjeeling Limited	2007	7	185127	Wes Anderson	Owen Wilson	Adrien Brody	Jason Schwartzman

In [241…
```
ax = sns.barplot(x='Votes', y=Top5_movies.index,data=Top5_movies, hue='Director')
for bars in ax. containers:
    ax.bar_label(bars)
```



Lower_rated movies

In [243…
```
df.nsmallest(5, columns=['Rating'], keep='first')
```

Out[243…

	Name	Year	Duration	Genre	Rating	Votes	Director	Actor 1	Actor 2	Actor 3
2918	Chatur Singh Two Star	2011	150	Action, Adventure, Comedy	1	624	Ajay Chandhok	Sanjay Dutt	Ameesha Patel	Anupam Kher
3618	Desh Drohi	2008	140	Action, Thriller	1	3899	Jagdish A. Sharma	Kamal Rashid Khan	Gracy Singh	Hrishitaa Bhatt
4170	Dracula 2012	2013	137	Horror	1	128	Vinayan	Thilakan	Shraddha Das	Monal Gajjar
5523	Hari Puttar: A Comedy of Terrors	2008	90	Comedy, Drama, Family	1	314	Rajesh Bajaj	Lucky Kohli	Jackie Shroff	Sarika
5711	Himmatwala	2013	150	Action, Comedy, Drama	1	8186	Sajid Khan	Ajay Devgn	Tamannaah Bhatia	Mahesh Manjrekar

Least_voted movies

In [284…
```
df.nsmallest(50, columns=['Votes'], keep='first')
```

Out[284…

	Name	Year	Duration	Genre	Rating	Votes	Director	Actor 1	Actor 2	Actor 3
1116	Anmol Sitaare	1982	153	Drama	7	5	Geethapriya	Master Baboo	Rakesh Bedi	Ramesh Deo
1341	Atal Faisla	2018	127	Drama	8	5	Abdul Sattar	Sahil Akhtar	Himayat Ali	Aman Jain
1434	Awesome Mausam	2016	131	Romance	3	5	Yogesh Bharadwaj	Vaarssh Bhatnagar	Sunil Chaurasiyaa	Suhasini Mulay
1469	B for Bundelkhand	2017	117	Drama	8	5	Vishal Mourya	Bharat Chawla	Nemi Chandra Jha	Moumita Nandi
2118	Bhai-Bahen	1959	136	Drama, Family	6	5	G.P. Sippy	Daisy Irani	Rajan Kapoor	Kathana
3061	Chipku	2016	135	Drama	6	5	Priyanka	Gouri Ash	Rakesh Bedi	Neha Pal Chowdhury
3107	Chori Chori	1972	115	Action, Drama	5	5	Keval P. Kashyap	Sanjay Khan	Radha Saluja	Kamran
3361	Daku Kali Bhawani	2000	120	Action	3	5	S.R. Pratap	Dharmendra	Milind Gunaji	Mohan Joshi
4046	Do Nambar Ke Amir	1974	136	Crime, Drama	7	5	P.D. Shenoy	Asha Sachdev	Sajid Khan	A.K. Hangal
4384	Ek Daku Saher Mein	1985	133	Action	3	5	Kalidas	Ashok Kumar	Pradeep Kumar	Suresh Oberoi
4523	Ek Tha Hero	2018	120	Drama, Family	4	5	Yogesh Pagare	Asrani	Ashwini Kalsekar	Ayush Mahesh Khedekar
5087	Ghulam Begum Badshah	1956	124	Drama, History	6	5	Jugal Kishore	Daljeet	Malika	Nishi
5353	Gyara Hazar Ladkian	1962	152	Drama	5	5	Khwaja Ahmad Abbas	Mala Sinha	Bharat Bhushan	Helen
6160	Intezar	1973	116	Drama	5	5	Mohan Singh Kavia	Rinku Jaiswal	Padmini Kapila	Baldev Khosa
6284	Jaan Lada Denge	1990	137	Action	6	5	Dilip Gulati	Hemant Birje	Sahila Chaddha	Natasha
6457	Jai Mahakali	1951	162	Fantasy	7	5	Dhirubhai Desai	Nirupa Roy	Shahu Modak	Ulhas
6573	Jawab Ayega	1968	82	Family	6	5	Ismat Chughtai	Shaheed Latif	Yogesh Bali	Meena Rai
7082	Kadke Kamal Ke	2019	136	Comedy	6	5	Laxman Singh	Aaryan Adhikari	Neeta Dhungana	Rajpal Yadav
7262	Kaneez	1949	140	Drama, History	6	5	Krishna Kumar	Munawwar Sultana	Shyam	Urmila
8288	Lorni - The Flaneur	2019	107	Crime, Mystery	7	5	Wanphrang Diengdoh	Poonam Agarwal	Elizer Bareh	Deihokhlang Basaiawmoit
8339	Love Qubool Hai	2020	94	Drama, Romance	10	5	Saif Ali Sayeed	Ahaan Jha	Mahesh Narayan	Rajasree Rajakumari
8726	Majaz: Ae Gham-e-Dil Kya Karun	2017	178	Biography	5	5	Ravindra Singh	Priyanshu Chatterjee	Rashmi Mishra	Neelima Azim
8801	Manchala	1999	137	Romance	5	5	Jayprakkash Shaw	Vivek Mushran	Gauri Khopkar	Rakesh Bedi
8923	Mashaal	1950	136	Drama	6	5	Nitin Bose	Ashok Kumar	Sumitra Devi	Ruma Guha Thakurta
9525	Mounto	1975	114	Action, Crime	4	5	Jambulingam	Navin Nischol	Saira Banu	Bindu
9568	Mr. X	1987	120	Fantasy	7	5	Khwaja Ahmad Abbas	Amol Palekar	Shabana Azmi	Tom Alter
9829	Naami Chor	1977	127	Action	5	5	Manmohan Desai	Kamal Mehra	Biswajit Chatterjee	Shatrughan Sinha
10179	Night Club	1958	137	Crime, Drama	6	5	Naresh Saigal	Ashok Kumar	Kamini Kaushal	Mubarak
11179	Pyaar Karle	2019	116	Drama	3	5	Rudra Patla Venugopal	Chitram Basha	Udita Goswami	Faheem Ul Haq
11352	Qatil Chudail	2001	96	Horror	5	5	Kanti Shah	Anil Nagrath	Jhony Nirmal	Amit Pachori
11628	Rakshaa Bandhan	1977	90	Drama	7	5	Shantilal Soni	Jairaj	Pallavi Joshi	Lalita Pawar
12058	Rustom-E-Rome	1964	106	Action, Adventure, Drama	6	5	Radhakant	Dara Singh	Azaad Irani	Vijaya Choudhury
12885	Shankar Khan	1966	140	Action, Drama, Sport	5	5	Nanabhai Bhatt	Prithviraj Kapoor	Dara Singh	Randhawa
12969	Shehnai	1947	133	Romance	4	5	P.L. Santoshi	Rehana	Indumati	Nasir Khan
13759	Tangewali	1955	120	Drama, Family	8	5	Lekhraj Bhakri	Shammi Kapoor	Balraj Sahni	Anita Guha
14825	Veeru Ustaad	1975	127	Drama	5	5	Jagdish Nirula	Aruna Irani	Shakti Kapoor	Satish Kaul
15466	Zindagi Aur Maut	1965	134	Drama	6	5	Nisar Ahmad Ansari	Nisar Ahmad Ansari	Chandrima Bhaduri	Bela Bose
79	3.0 Megapixel	2015	99	Adventure, Drama	4	6	Divakar Ghodake	Seema Azmi	Meenal Chirankar	Pooja Kasekar
91	3rd EYE	2019	145	Horror	6	6	Sanjay Niranjan	Deepak Bhatia	Mukul Dev	Sakshi Dwivedi
493	Aasmaan Mahal	1965	172	Drama	5	6	Khwaja Ahmad Abbas	Prithviraj Kapoor	Dilip Roy	Surekha
530	Ab Hoga Dharna Unlimited	2012	78	Comedy	4	6	Deepak Tanwar	Saurabh Malik	Rekha Rana	Omkar Das Manikpuri
539	Ab To Aaja Saajan Mere	1994	127	Drama, Romance	4	6	Rakesh Nahata	Keshto Iqbal	Arvind Joshi	Shalini Kapoor
1189	Apaatkaal	1993	137	Action	5	6	V. Subba Rao	Aatish Devgan	Sabah	Rita Bhaduri
1236	Apradhi	1974	122	Drama	7	6	Jugal Kishore	Kiran Kumar	Yogeeta Bali	Imtiaz
1531	Baazigar	1972	133	Adventure, Comedy, Crime	7	6	Karunesh Thakur	Vijayalalitha	Roopesh Kumar	Bindu
1575	Bad Boys	2003	118	Thriller	3	6	Vicky S. Kumar	Rakesh Bedi	Suresh Chatwal	Avtar Gill
1684	Bagpat Ka Dulha	2019	137	Drama	5	6	Karan Kashyap	Raza Murad	Puneet Vashisht	Lalit Parimoo
1861	Banwra	1950	148	Drama	6	6	G. Rakesh	Raj Kapoor	Lalita Pawar	Nimmi
1932	Bazaar	1949	148	Drama, Family	6	6	K. Amarnath	Nigar Sultana	Shyam	Gope
2207	Bhayaanak	1998	105	Horror	4	6	R. Mittal	Anil Dhawan	Tina Ghai	Bharat Kapoor

Insights

1. Out of 10 the 6 ratings are mostly given to movies at a maximum of 1661 counts.

2. Gangs of Wasseypur have taken a maximum duration, of 321 minutes. Action, comedy, and crime movie made in 2012 by Anurag Kashyap director.

3. The highest voting goes to life of pi movie, 591417. Life of pi was released in 2012 by Ang lee director.

4. Least_voted movies are like, Anmol sitaare, Atal Faisla, Awesome Mausam (Flop), B for Bundelkhand, Bhai-behan, and many more. Overall 37 movies got the least voting.

5. There is one movie, Love Qubool Hai got 10 out of 10 rating (highest_rated) directed by Saif Ali Sayeed but got the least number of votes.