Movie rating prediction

The aim of analysis in this project is to predict movie rating based on factors like year, genre, directors, and actors. Detailed Exploratory Data Analysis and Machine learning algorithms have been used to predict ratings of movies.

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

In [2]:
```
df = pd.read_csv('IMDb Movies India.csv', encoding='latin1')
df.head(20)
```

Out[2]:

	Name	Year	Duration	Genre	Rating	Votes	Director	Actor 1	Actor 2
0		NaN	NaN	Drama	NaN	NaN	J.S. Randhawa	Manmauji	Birba
1	#Gadhvi (He thought he was Gandhi)	(2019)	109 min	Drama	7.0	8	Gaurav Bakshi	Rasika Dugal	Vive Ghamande
2	#Homecoming	(2021)	90 min	Drama, Musical	NaN	NaN	Soumyajit Majumdar	Sayani Gupta	Plabita Borthaku
3	#Yaaram	(2019)	110 min	Comedy, Romance	4.4	35	Ovais Khan	Prateik	Ishita Ra
4	...And Once Again	(2010)	105 min	Drama	NaN	NaN	Amol Palekar	Rajat Kapoor	Rituparna Sengupta
5	...Aur Pyaar Ho Gaya	(1997)	147 min	Comedy, Drama, Musical	4.7	827	Rahul Rawail	Bobby Deol	Aishwarya Ra Bachchan
6	...Yahaan	(2005)	142 min	Drama, Romance, War	7.4	1,086	Shoojit Sircar	Jimmy Sheirgill	Minissha Lamba
7	.in for Motion	(2008)	59 min	Documentary	NaN	NaN	Anirban Datta	NaN	NaN
8	?: A Question Mark	(2012)	82 min	Horror, Mystery, Thriller	5.6	326	Allyson Patel	Yash Dave	Muntazir Ahmad
9	@Andheri	(2014)	116 min	Action, Crime, Thriller	4.0	11	Biju Bhaskar Nair	Augustine	Fathima Babu
10	1:1.6 An Ode to Lost Love	(2004)	96 min	Drama	6.2	17	Madhu Ambat	Rati Agnihotri	Gulshan Grover
11	1:13:7 Ek Tera Saath	(2016)	120 min	Horror	5.9	59	Arshad Siddiqui	Pankaj Berry	Anubhav Dhi
12	100 Days	(1991)	161 min	Horror, Romance, Thriller	6.5	983	Partho Ghosh	Jackie Shroff	Madhuri Dixi
13	100% Love	(2012)	166 min	Comedy, Drama, Romance	5.7	512	Rabi Kinagi	Jeet	Koyel Mallic
14	101 Ratein	(1990)	NaN	Thriller	NaN	NaN	Harish	Saraswati	Disco Shant
15	102 Not Out	(2018)	102 min	Comedy, Drama	7.4	6,619	Umesh Shukla	Amitabh Bachchan	Rish Kapoo
16	108 Limited	NaN	NaN	NaN	NaN	NaN	Anand Anddy	Vijay Raaz	Sanjay Mishra
17	108 Teerthyatra	(1987)	NaN	Comedy, Drama, Fantasy	NaN	NaN	Rajpati	Pravin Anand	Nayan Bhat
18	10ml LOVE	(2010)	87 min	Comedy, Drama, Romance	6.3	162	Sharat Katariya	Neil Bhoopalam	Anusha Bos
19	11 O'Clock	(1948)	NaN	NaN	NaN	NaN	Homi	Aftab	Sayan

	Name	Year	Duration	Genre	Rating	Votes	Director	Actor 1	Actor 2

In [3]:
```
df.tail(5)
```

Out[3]:

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

In [4]:
```
df.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
```

In [6]:
```
df.describe()
```

Out[6]:

	Rating
count	7919.000000
mean	5.841621
std	1.381777
min	1.100000
25%	4.900000
50%	6.000000
75%	6.800000
max	10.000000

In [8]:
```
print("Number of rows:",df.shape[0])
print("Number of columns:",df.shape[1])
```

```
Number of rows: 15509
Number of columns: 10
```

In [9]:
```
df.isnull().sum()
```

Out[9]:
```
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 [12]:
```
missing_precentage=df.isnull().mean()*100
print(missing_precentage)
```

```
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
```

In [13]:
```
#dropping null values for %<10
df.dropna(subset=['Year'],inplace=True)
```

In [14]:
```
df.head(8)
```

Out[14]:

	Name	Year	Duration	Genre	Rating	Votes	Director	Actor 1	Actor 2	A
1	#Gadhvi (He thought he was Gandhi)	(2019)	109 min	Drama	7.0	8	Gaurav Bakshi	Rasika Dugal	Vivek Ghamande	
2	#Homecoming	(2021)	90 min	Drama, Musical	NaN	NaN	Soumyajit Majumdar	Sayani Gupta	Plabita Borthakur	
3	#Yaaram	(2019)	110 min	Comedy, Romance	4.4	35	Ovais Khan	Prateik	Ishita Raj	Si
4	...And Once Again	(2010)	105 min	Drama	NaN	NaN	Amol Palekar	Rajat Kapoor	Rituparna Sengupta	
5	...Aur Pyaar Ho Gaya	(1997)	147 min	Comedy, Drama, Musical	4.7	827	Rahul Rawail	Bobby Deol	Aishwarya Rai Bachchan	S
6	...Yahaan	(2005)	142 min	Drama, Romance, War	7.4	1,086	Shoojit Sircar	Jimmy Sheirgill	Minissha Lamba	
7	.in for Motion	(2008)	59 min	Documentary	NaN	NaN	Anirban Datta	NaN	NaN	
8	?: A Question Mark	(2012)	82 min	Horror, Mystery, Thriller	5.6	326	Allyson Patel	Yash Dave	Muntazir Ahmad	

In [15]:
```
missing_precentage=df.isnull().mean()*100
print(missing_precentage)
```

```
Name          0.000000
Year          0.000000
Duration     52.506508
Genre        12.288899
Rating       47.139710
Votes        47.133035
Director      3.317536
Actor 1       9.932581
Actor 2      14.665243
Actor 3      19.404579
dtype: float64
```

In [19]:
```
df.dropna(subset=['Actor 1'],inplace=True)
```

In [20]:
```
df.dropna(subset=['Director'],inplace=True)
```

In [21]:
```
missing_precentage=df.isnull().mean()*100
print(missing_precentage)
```

```
Name          0.000000
Year          0.000000
Duration     49.099533
Genre        10.850070
Rating       42.236715
Votes        42.229304
Director      0.000000
Actor 1       0.000000
Actor 2       5.254576
Actor 3      10.516564
dtype: float64
```

In [23]:
```
df.dropna(subset=['Actor 2'],inplace=True)
```

In [24]:
```
df.dropna(subset=['Actor 3'],inplace=True)
```

In [25]:
```
missing_precentage=df.isnull().mean()*100
print(missing_precentage)
```

```
Name         0.000000
Year         0.000000
Duration    45.279112
Genre        8.182872
Rating      36.831208
Votes       36.822925
Director     0.000000
Actor 1      0.000000
Actor 2      0.000000
Actor 3      0.000000
dtype: float64
```

In [27]:
```
df.dropna(subset=['Genre'],inplace=True)
missing_precentage=df.isnull().mean()*100
print(missing_precentage)
```

```
Name         0.000000
Year         0.000000
Duration    41.863612
Genre        0.000000
Rating      31.823922
Votes       31.814902
Director     0.000000
Actor 1      0.000000
Actor 2      0.000000
Actor 3      0.000000
dtype: float64
```

In [28]:
```
#filling missing values of Duration by mean
df['Duration'] = df['Duration'].str.replace('min', '')
df['Duration'] = pd.to_numeric(df['Duration'], errors='coerce')
df['Duration'].fillna(df['Duration'].mean(), inplace=True)
df['Duration'] = df['Duration'].astype(int)
df.head(8)
```

Out[28]:

	Name	Year	Duration	Genre	Rating	Votes	Director	Actor 1	Actor 2	Ac
1	#Gadhvi (He thought he was Gandhi)	(2019)	109	Drama	7.0	8	Gaurav Bakshi	Rasika Dugal	Vivek Ghamande	A J
2	#Homecoming	(2021)	90	Drama, Musical	NaN	NaN	Soumyajit Majumdar	Sayani Gupta	Plabita Borthakur	An
3	#Yaaram	(2019)	110	Comedy, Romance	4.4	35	Ovais Khan	Prateik	Ishita Raj	Sid Ka
4	...And Once Again	(2010)	105	Drama	NaN	NaN	Amol Palekar	Rajat Kapoor	Rituparna Sengupta	A
5	...Aur Pyaar Ho Gaya	(1997)	147	Comedy, Drama, Musical	4.7	827	Rahul Rawail	Bobby Deol	Aishwarya Rai Bachchan	Sh Ka
6	...Yahaan	(2005)	142	Drama, Romance, War	7.4	1,086	Shoojit Sircar	Jimmy Sheirgill	Minissha Lamba	Ya Sh
8	?: A Question Mark	(2012)	82	Horror, Mystery, Thriller	5.6	326	Allyson Patel	Yash Dave	Muntazir Ahmad	E
9	@Andheri	(2014)	116	Action, Crime, Thriller	4.0	11	Biju Bhaskar Nair	Augustine	Fathima Babu	

In [29]:
```
missing_precentage=df.isnull().mean()*100
print(missing_precentage)
```

```
Name          0.000000
Year          0.000000
Duration      0.000000
Genre         0.000000
Rating       31.823922
Votes        31.814902
Director      0.000000
Actor 1       0.000000
Actor 2       0.000000
Actor 3       0.000000
dtype: float64
```

In [30]:
```
#filling missing valuse of Votes by mean
df['Votes'] = pd.to_numeric(df['Votes'], errors='coerce')
df['Votes'].fillna(df['Votes'].mean(), inplace=True)
df['Votes'] = df['Votes'].astype(int)
df.head(9)
```

Out[30]:

	Name	Year	Duration	Genre	Rating	Votes	Director	Actor 1	Actor 2	A
1	#Gadhvi (He thought he was Gandhi)	(2019)	109	Drama	7.0	8	Gaurav Bakshi	Rasika Dugal	Vivek Ghamande	
2	#Homecoming	(2021)	90	Drama, Musical	NaN	125	Soumyajit Majumdar	Sayani Gupta	Plabita Borthakur	A
3	#Yaaram	(2019)	110	Comedy, Romance	4.4	35	Ovais Khan	Prateik	Ishita Raj	Si
4	...And Once Again	(2010)	105	Drama	NaN	125	Amol Palekar	Rajat Kapoor	Rituparna Sengupta	
5	...Aur Pyaar Ho Gaya	(1997)	147	Comedy, Drama, Musical	4.7	827	Rahul Rawail	Bobby Deol	Aishwarya Rai Bachchan	S
6	...Yahaan	(2005)	142	Drama, Romance, War	7.4	125	Shoojit Sircar	Jimmy Sheirgill	Minissha Lamba	Y S
8	?: A Question Mark	(2012)	82	Horror, Mystery, Thriller	5.6	326	Allyson Patel	Yash Dave	Muntazir Ahmad	
9	@Andheri	(2014)	116	Action, Crime, Thriller	4.0	11	Biju Bhaskar Nair	Augustine	Fathima Babu	
10	1:1.6 An Ode to Lost Love	(2004)	96	Drama	6.2	17	Madhu Ambat	Rati Agnihotri	Gulshan Grover	K

In [31]:
```
missing_precentage=df.isnull().mean()*100
print(missing_precentage)
```

```
Name          0.000000
Year          0.000000
Duration      0.000000
Genre         0.000000
Rating       31.823922
Votes         0.000000
Director      0.000000
Actor 1       0.000000
Actor 2       0.000000
Actor 3       0.000000
dtype: float64
```

In [32]:
```
#dropping missing values from ratings
df.dropna(subset=['Rating'],inplace=True)
```

In [33]:
```
df.shape
```

Out[33]:
```
(7558, 10)
```

DATA VISUALISATION

In [34]:
```
df.duplicated().any()
```

Out[34]:
```
False
```

year with highest number of votes

In [39]:
```
df.groupby('Year')['Votes'].mean().sort_values(ascending=False)
```

Out[39]:
```
Year
(2013)    191.055556
(2003)    182.773973
(2008)    181.785185
(2010)    179.842105
(2009)    178.130719
             ...
(1944)     13.454545
(1938)      9.714286
(1932)      9.000000
(1934)      8.500000
(1939)      8.250000
Name: Votes, Length: 92, dtype: float64
```

Highest rated genre

In [40]:
```
genre=df.groupby('Genre')['Rating'].mean().sort_values(ascending=False)
```

Out[40]:
```
Genre
History, Romance                9.4
Documentary, Family, History    9.3
Documentary, Music              8.9
Documentary, Thriller           8.7
Documentary, Sport              8.6
                               ...
Action, Fantasy, Sci-Fi         2.7
Comedy, Horror, Musical         2.7
Family, Music, Romance          2.6
Action, Comedy, Horror          2.4
Comedy, Family, Sci-Fi          2.4
Name: Rating, Length: 416, dtype: float64
```

Highest rated directors

In [45]:
```
df.groupby('Director')['Rating'].mean().sort_values(ascending=False)
```

Out[45]:
```
Director
Saif Ali Sayeed    10.0
Sriram Raja         9.7
Bobby Kumar         9.6
Arvind Pratap       9.4
Munni Pankaj        9.4
                   ...
Umesh Ghadge        1.9
Rajesh Bajaj        1.9
Stanley D'Costa     1.8
Raajeev Walia       1.8
Pramod Mandloi      1.7
Name: Rating, Length: 2956, dtype: float64
```

Top 10 longest duration movies

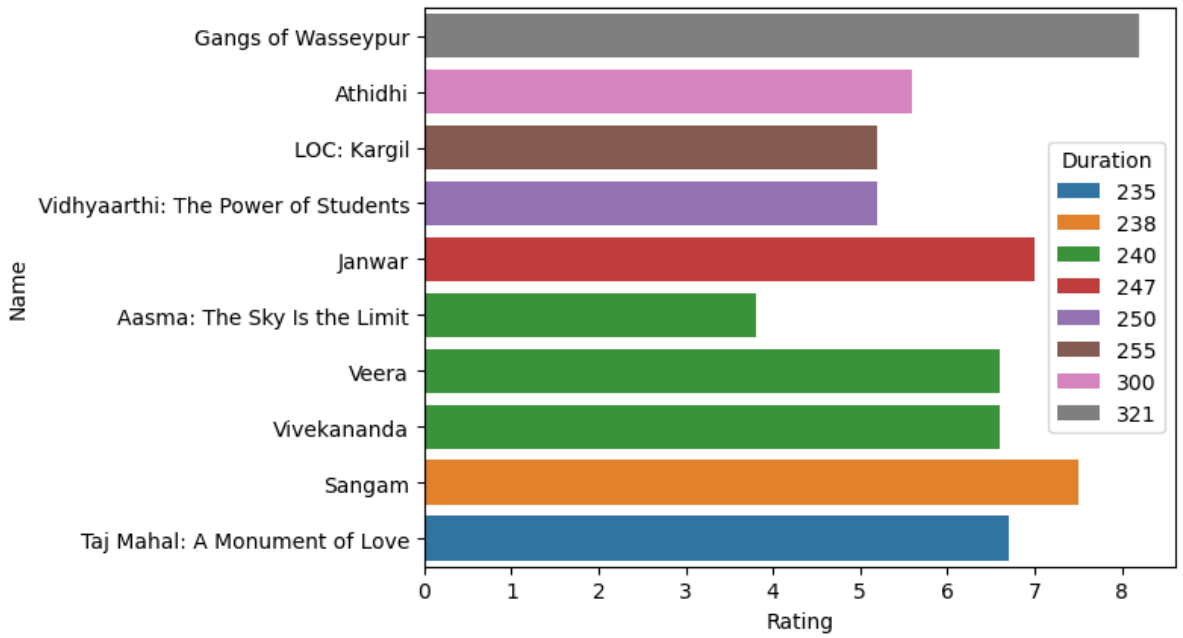
In [51]:
```
high_duration=df.nlargest(10,'Duration')[['Name','Duration','Rating']].set_index('N
high_duration
```

Out[51]:

Name	Duration	Rating
Gangs of Wasseypur	321	8.2
Athidhi	300	5.6
LOC: Kargil	255	5.2
Vidhyaarthi: The Power of Students	250	5.2
Janwar	247	7.0
Aasma: The Sky Is the Limit	240	3.8
Veera	240	6.6
Vivekananda	240	6.6
Sangam	238	7.5
Taj Mahal: A Monument of Love	235	6.7

In [54]:
```
sns.barplot(x='Rating',y=high_duration.index,data=high_duration,hue='Duration',dodg
```

Out[54]:
```
<Axes: xlabel='Rating', ylabel='Name'>
```



Top 10 directors

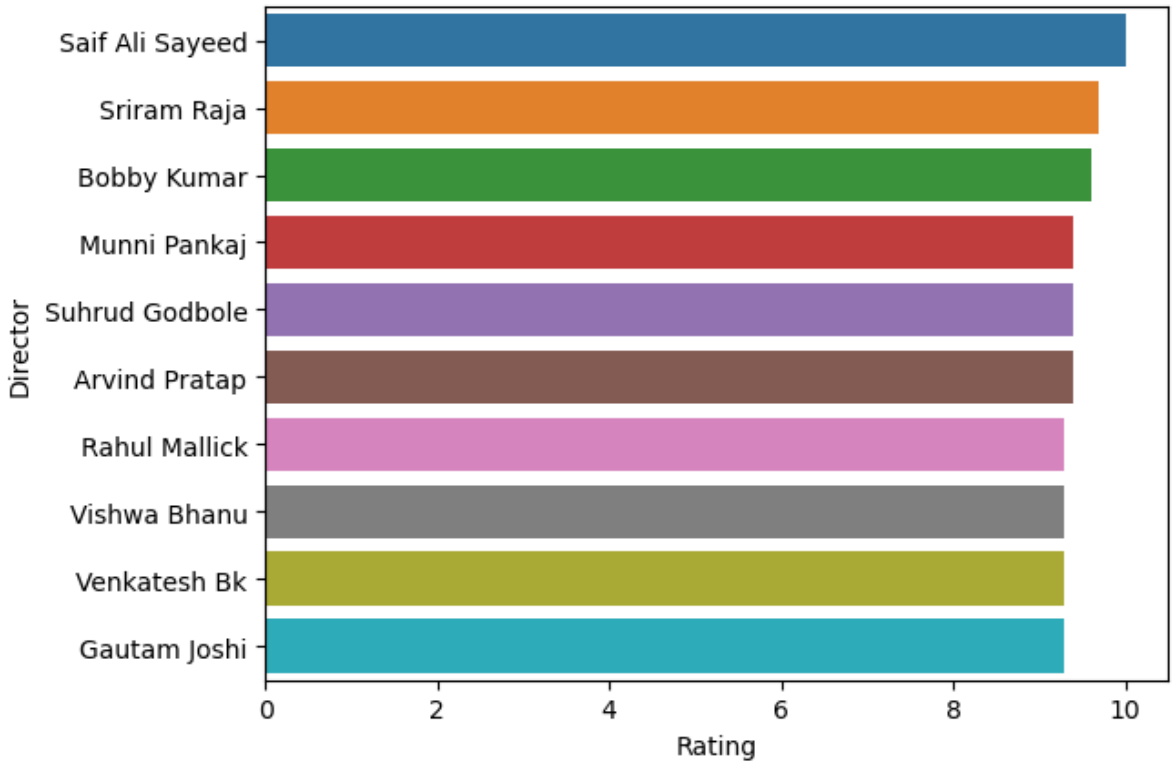
In [56]:
```
high_direc=df.nlargest(10,'Rating')[['Director','Rating']].set_index('Director')
high_direc
```

Out[56]:

Director	Rating
Saif Ali Sayeed	10.0
Sriram Raja	9.7
Bobby Kumar	9.6
Munni Pankaj	9.4
Suhrud Godbole	9.4
Arvind Pratap	9.4
Rahul Mallick	9.3
Vishwa Bhanu	9.3
Venkatesh Bk	9.3
Gautam Joshi	9.3

In [57]:
```
sns.barplot(x='Rating',y=high_direc.index,data=high_direc)
```

Out[57]:
```
<Axes: xlabel='Rating', ylabel='Director'>
```



top 10 actors

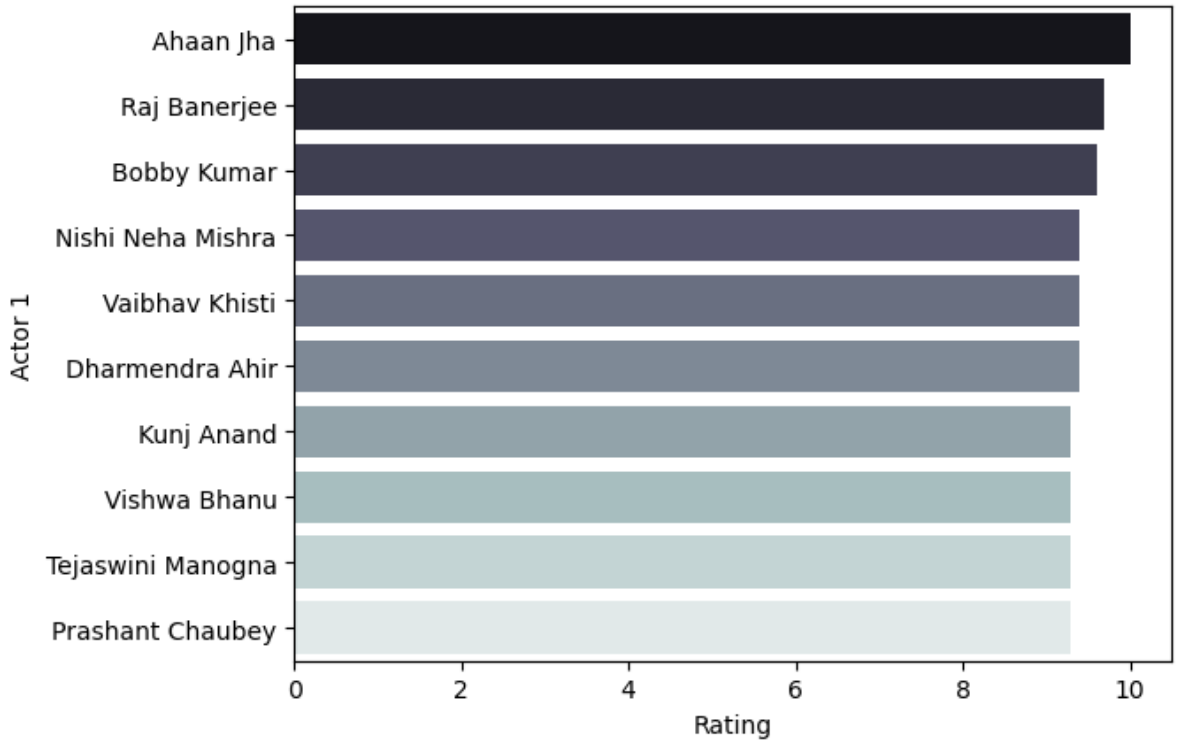
In [58]:
```
top_actor=df.nlargest(10,'Rating')[['Actor 1','Rating']].set_index('Actor 1')
top_actor
```

Out[58]:

Actor 1	Rating
Ahaan Jha	10.0
Raj Banerjee	9.7
Bobby Kumar	9.6
Nishi Neha Mishra	9.4
Vaibhav Khisti	9.4
Dharmendra Ahir	9.4
Kunj Anand	9.3
Vishwa Bhanu	9.3
Tejaswini Manogna	9.3
Prashant Chaubey	9.3

In [62]:
```
sns.barplot(x='Rating',y=top_actor.index,data=top_actor,palette="bone")
```

Out[62]:
```
<Axes: xlabel='Rating', ylabel='Actor 1'>
```



Feature engineering

In [120…
```
df.columns
```

Out[120]:
```
Index(['Name', 'Year', 'Duration', 'Genre', 'Rating', 'Votes', 'Director',
       'Actor 1', 'Actor 2', 'Actor 3'],
      dtype='object')
```

In [121…
```
dp=df.drop(['Name'],axis=1)
dp.head(8)
```

Out[121]:

	Year	Duration	Genre	Rating	Votes	Director	Actor 1	Actor 2	Actor 3
1	(2019)	109	Drama	7.0	8	Gaurav Bakshi	Rasika Dugal	Vivek Ghamande	Arvind Jangid
3	(2019)	110	Comedy, Romance	4.4	35	Ovais Khan	Prateik	Ishita Raj	Siddhant Kapoor
5	(1997)	147	Comedy, Drama, Musical	4.7	827	Rahul Rawail	Bobby Deol	Aishwarya Rai Bachchan	Shammi Kapoor
6	(2005)	142	Drama, Romance, War	7.4	125	Shoojit Sircar	Jimmy Sheirgill	Minissha Lamba	Yashpal Sharma
8	(2012)	82	Horror, Mystery, Thriller	5.6	326	Allyson Patel	Yash Dave	Muntazir Ahmad	Kiran Bhatia
9	(2014)	116	Action, Crime, Thriller	4.0	11	Biju Bhaskar Nair	Augustine	Fathima Babu	Byon
10	(2004)	96	Drama	6.2	17	Madhu Ambat	Rati Agnihotri	Gulshan Grover	Atul Kulkarni
11	(2016)	120	Horror	5.9	59	Arshad Siddiqui	Pankaj Berry	Anubhav Dhir	Hritu Dudani

In [122…

```
actor1=dp.groupby('Actor 1').agg({'Rating':'mean'}).to_dict()
actor2=dp.groupby('Actor 2').agg({'Rating':'mean'}).to_dict()
actor3=dp.groupby('Actor 3').agg({'Rating':'mean'}).to_dict()
genre=dp.groupby('Genre').agg({'Rating':'mean'}).to_dict()
director=dp.groupby('Director').agg({'Rating':'mean'}).to_dict()
```

In [123…

```
dp['actor1'] = round(dp['Actor 1'].map(actor1['Rating']),1)
dp['actor2'] = round(dp['Actor 2'].map(actor2['Rating']),1)
dp['actor3'] = round(dp['Actor 3'].map(actor3['Rating']),1)
dp['director'] = round(dp['Director'].map(director['Rating']),1)
dp['genre'] = round(dp['Genre'].map(genre['Rating']),1)
```

In [124…

```
dp.head(7)
```

Out[124]:

	Year	Duration	Genre	Rating	Votes	Director	Actor 1	Actor 2	Actor 3	actor1	a
1	(2019)	109	Drama	7.0	8	Gaurav Bakshi	Rasika Dugal	Vivek Ghamande	Arvind Jangid	6.8	
3	(2019)	110	Comedy, Romance	4.4	35	Ovais Khan	Prateik	Ishita Raj	Siddhant Kapoor	5.4	
5	(1997)	147	Comedy, Drama, Musical	4.7	827	Rahul Rawail	Bobby Deol	Aishwarya Rai Bachchan	Shammi Kapoor	4.8	
6	(2005)	142	Drama, Romance, War	7.4	125	Shoojit Sircar	Jimmy Sheirgill	Minissha Lamba	Yashpal Sharma	5.3	
8	(2012)	82	Horror, Mystery, Thriller	5.6	326	Allyson Patel	Yash Dave	Muntazir Ahmad	Kiran Bhatia	5.6	
9	(2014)	116	Action, Crime, Thriller	4.0	11	Biju Bhaskar Nair	Augustine	Fathima Babu	Byon	4.0	
10	(2004)	96	Drama	6.2	17	Madhu Ambat	Rati Agnihotri	Gulshan Grover	Atul Kulkarni	5.2	

In [125…

```
dp.drop(columns=['Actor 1','Actor 2','Actor 3','Genre'],axis=1,inplace=True)
dp.head(8)
```

Out[125]:

	Year	Duration	Rating	Votes	Director	actor1	actor2	actor3	director	genre
1	(2019)	109	7.0	8	Gaurav Bakshi	6.8	7.0	7.0	7.0	6.3
3	(2019)	110	4.4	35	Ovais Khan	5.4	4.4	4.4	4.4	5.7
5	(1997)	147	4.7	827	Rahul Rawail	4.8	5.8	5.8	5.4	6.2
6	(2005)	142	7.4	125	Shoojit Sircar	5.3	6.0	6.5	7.5	6.8
8	(2012)	82	5.6	326	Allyson Patel	5.6	5.9	5.6	5.6	5.5
9	(2014)	116	4.0	11	Biju Bhaskar Nair	4.0	4.6	4.0	4.0	5.3
10	(2004)	96	6.2	17	Madhu Ambat	5.2	5.4	5.2	6.2	6.3
11	(2016)	120	5.9	59	Arshad Siddiqui	5.8	5.9	5.9	7.0	4.6

In [126…

```
dp.drop(['Director'],axis=1,inplace=True)
dp.head(9)
```

Out[126]:

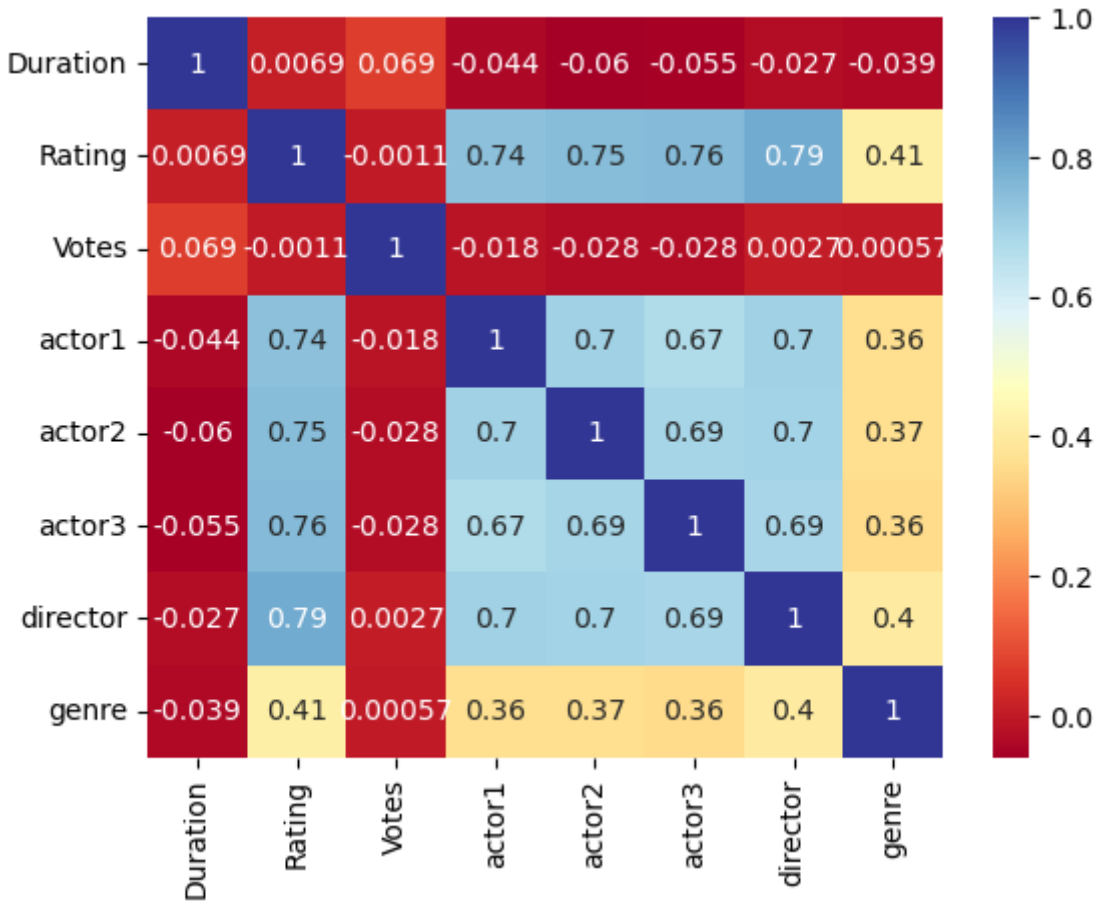
	Year	Duration	Rating	Votes	actor1	actor2	actor3	director	genre
1	(2019)	109	7.0	8	6.8	7.0	7.0	7.0	6.3
3	(2019)	110	4.4	35	5.4	4.4	4.4	4.4	5.7
5	(1997)	147	4.7	827	4.8	5.8	5.8	5.4	6.2
6	(2005)	142	7.4	125	5.3	6.0	6.5	7.5	6.8
8	(2012)	82	5.6	326	5.6	5.9	5.6	5.6	5.5
9	(2014)	116	4.0	11	4.0	4.6	4.0	4.0	5.3
10	(2004)	96	6.2	17	5.2	5.4	5.2	6.2	6.3
11	(2016)	120	5.9	59	5.8	5.9	5.9	7.0	4.6
12	(1991)	161	6.5	983	5.1	5.8	5.2	4.8	5.4

Model selection

In [127…

```
sns.heatmap(dp.corr(),annot=True,cmap='RdYlBu')
plt.show()
```

```
C:\Users\mairah nisar\AppData\Local\Temp\ipykernel_43952\2482915387.py:1: FutureWa
rning: The default value of numeric_only in DataFrame.corr is deprecated. In a fut
ure version, it will default to False. Select only valid columns or specify the va
lue of numeric_only to silence this warning.
  sns.heatmap(dp.corr(),annot=True,cmap='RdYlBu')
```



In [128…

```
from sklearn import datasets  # Import datasets
from sklearn.model_selection import train_test_split  # Import train_test_split fur
from sklearn.linear_model import LinearRegression  # Import LinearRegression class
from sklearn.tree import DecisionTreeClassifier  # Import DecisionTreeClassifier cl
from sklearn.cluster import KMeans
```

In [129…
```
X = dp.drop('Rating', axis=1)
y = dp['Rating']

X.head()
```

Out[129]:

	Year	Duration	Votes	actor1	actor2	actor3	director	genre
1	(2019)	109	8	6.8	7.0	7.0	7.0	6.3
3	(2019)	110	35	5.4	4.4	4.4	4.4	5.7
5	(1997)	147	827	4.8	5.8	5.8	5.4	6.2
6	(2005)	142	125	5.3	6.0	6.5	7.5	6.8
8	(2012)	82	326	5.6	5.9	5.6	5.6	5.5

In [131…
```
x_train, x_test,y_train,y_test = train_test_split(X,y,test_size =0.2)
# print the data
x_train
```

Out[131]:

	Year	Duration	Votes	actor1	actor2	actor3	director	genre
8120	(2002)	131	88	6.4	5.8	6.2	6.7	5.2
1634	(1998)	119	25	5.1	4.5	5.3	2.5	5.0
14697	(2004)	134	42	3.5	4.8	5.4	5.1	5.7
4563	(2012)	134	125	5.6	6.4	7.8	7.6	6.1
11436	(1970)	131	5	6.2	6.3	5.3	4.4	6.3
…	…	…	…	…	…	…	…	…
14989	(1982)	131	5	4.0	5.5	6.7	3.4	5.5
10888	(2018)	110	22	5.4	5.4	5.4	5.4	6.3
4125	(1983)	153	15	6.6	6.9	6.2	6.4	6.2
9998	(1978)	140	13	6.5	6.5	6.5	6.2	5.9
736	(1986)	131	36	5.3	5.5	5.6	6.2	5.9

6046 rows × 8 columns

In [139…
```
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.4, random_sta
```

In [140…
```
print('Shape of training features:', X_train.shape)
print('shape of training target:', y_train.shape)
print('Shape of testing features:', X_test.shape)
print('shape of testing target:', y_test.shape)
```

```
Shape of training features: (4534, 8)
shape of training target: (4534,)
Shape of testing features: (3024, 8)
shape of testing target: (3024,)
```

In [141…
```
regression=LinearRegression()
regression.fit(X_train,y_train)
```

Out[141]:

```
LinearRegression()
```

In [142…

```
print("coefficients:",regression.score(X_train,y_train))
y_pred_regression=regression.predict(X_test)
```

```
coefficients: 0.7603933925982155
```

In [143…

```
results= pd.DataFrame({'actual':y_test,'predicted':y_pred_regression.ravel(),'resid
results.head(50)
```

Out[143]:

	actual	predicted	residual
6241	7.4	6.651753	0.748247
3321	4.9	5.467040	-0.567040
6117	6.5	6.310843	0.189157
5975	5.7	5.553142	0.146858
6653	7.0	6.938941	0.061059
9928	4.1	4.553361	-0.453361
7846	5.7	5.104540	0.595460
14588	6.9	6.872685	0.027315
5735	5.4	5.434157	-0.034157
3483	6.9	7.119600	-0.219600
11376	5.8	5.599471	0.200529
7553	6.2	5.813279	0.386721
3040	6.7	7.028623	-0.328623
5367	6.2	5.426156	0.773844
3634	6.2	6.424889	-0.224889
11539	5.3	5.117312	0.182688
12487	8.1	6.275347	1.824653
2289	5.4	5.385914	0.014086
13623	6.5	5.683924	0.816076
5202	5.1	5.049009	0.050991
39	4.1	4.285893	-0.185893
14430	6.7	6.102853	0.597147
11145	4.6	5.050496	-0.450496
10897	6.2	6.309463	-0.109463
2358	3.8	3.917587	-0.117587
7058	6.8	6.387415	0.412585
6841	7.5	7.006655	0.493345
2593	6.6	6.488575	0.111425
1865	6.0	5.898915	0.101085
4215	6.2	6.427384	-0.227384
11981	3.5	2.659470	0.840530
3771	3.4	4.258263	-0.858263
9109	5.7	6.334497	-0.634497
9404	7.1	6.770331	0.329669
904	4.4	3.754614	0.645386
541	4.6	5.858579	-1.258579

	actual	predicted	residual
15494	6.2	6.238343	-0.038343
12835	4.9	4.676292	0.223708
3464	7.1	6.603569	0.496431
8182	3.7	3.678640	0.021360
7523	2.8	2.028341	0.771659
9389	4.3	4.108125	0.191875
5469	7.6	8.200632	-0.600632
15029	3.8	3.588587	0.211413
7720	7.1	7.066716	0.033284
13433	5.0	4.762630	0.237370
8169	5.6	5.540231	0.059769
5858	4.4	5.349623	-0.949623
3121	6.1	6.637125	-0.537125
2697	5.8	4.284183	1.515817

In [144...
```
 r_sq = regression.score(X, y)
print(f"coefficient of determination: {r_sq}")
```

```
coefficient of determination: 0.7589545932592199
```

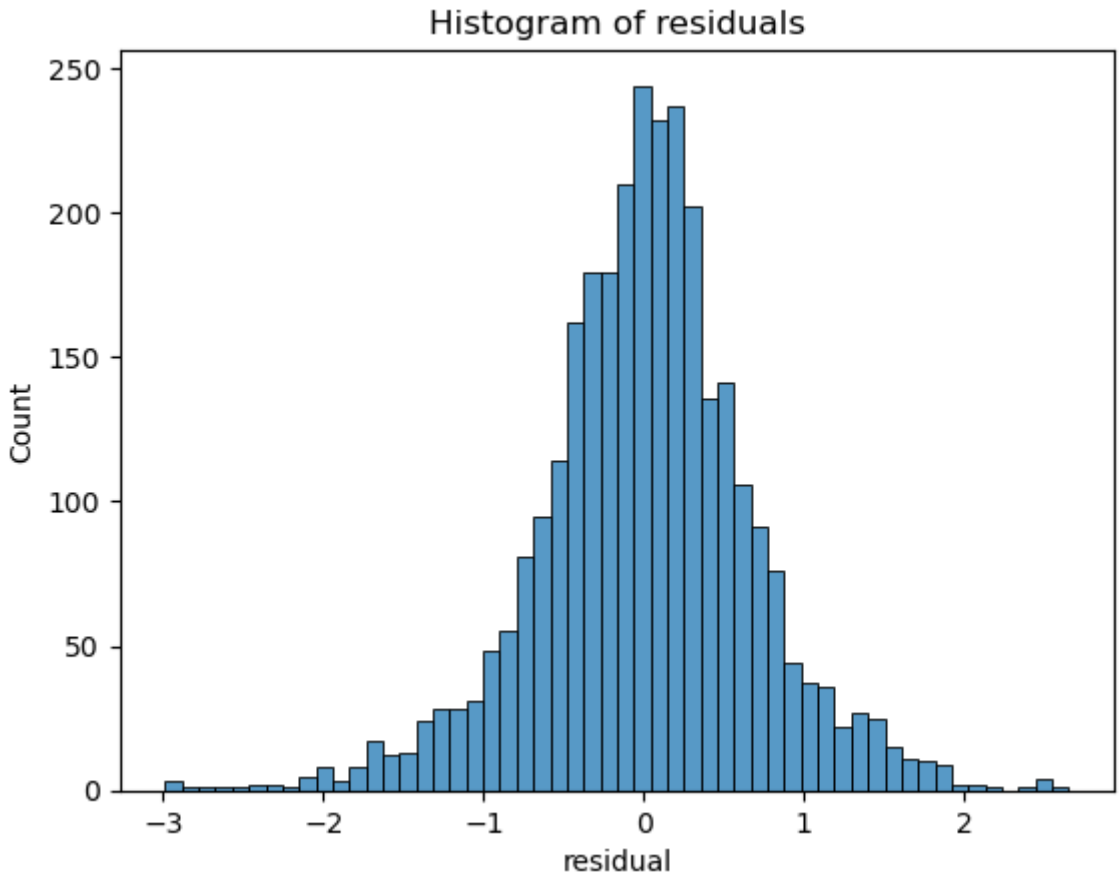
In [145...
```
print(f"intercept: {regression.intercept_}")
print(f"slope: {regression.coef_}")
```

```
intercept: -6.614685082409624
slope: [1.86781643e-03 3.76522748e-03 6.79622283e-05 2.58581986e-01
 2.63088359e-01 3.40023280e-01 4.09857955e-01 1.38288004e-01]
```

In [146...
```
from sklearn.model_selection import train_test_split, cross_val_score, KFold, GridS

k = 5 # Number of folds
cv = KFold(n_splits=k, shuffle=True, random_state=42)

scores = cross_val_score(regression, X, y, cv=cv, scoring='r2')


print("R^2 scores:", scores)
print("Mean R^2:", scores.mean())
print("Standard Deviation of R^2:", scores.std())
```

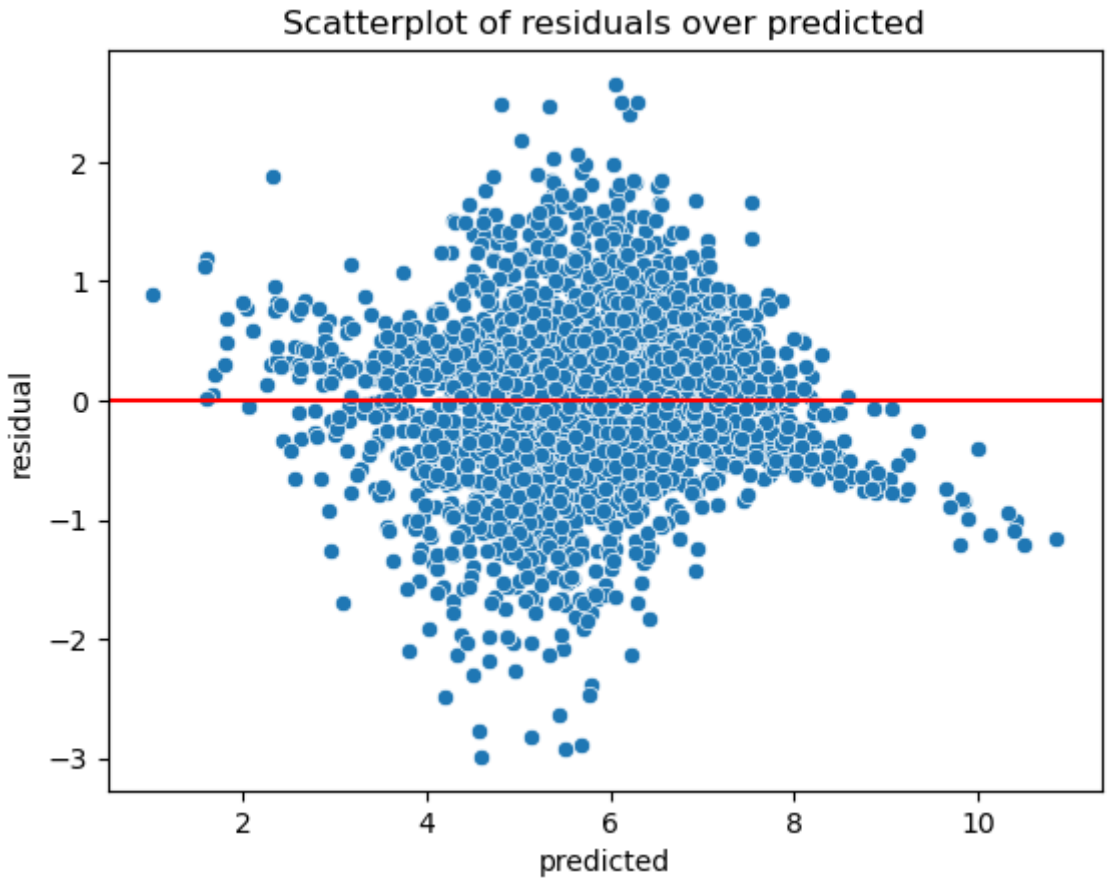
```
R^2 scores: [0.7548157  0.75858315 0.76045957 0.75852977 0.75995955]
Mean R^2: 0.7584695483808084
Standard Deviation of R^2: 0.0019772401648770455
```

In [148...
```
sns.histplot(results['residual'])
plt.title('Histogram of residuals')
plt.show()
```



In [149…

```
sns.scatterplot(x=results['predicted'], y=results['residual'])
plt.axhline(0, c='red')
plt.title('Scatterplot of residuals over predicted')
plt.show()
```



In [ ]:

In [ ]:

In [ ]:

In [ ]:

In [ ]:

In [ ]: