In [106]:

```
import pandas as pd
import numpy as np
import nltk
from nltk.sentiment.vader import SentimentIntensityAnalyzer
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LinearRegression
from sklearn.metrics import r2_score,mean_absolute_error
```

In [70]:

```
nltk.download('vader_lexicon')
```

```
[nltk_data] Downloading package vader_lexicon to
[nltk_data]     C:\Users\ssneh\AppData\Roaming\nltk_data...
[nltk_data]   Package vader_lexicon is already up-to-date!
```

Out[70]:

```
True
```

In [72]:

```
df=pd.read_csv("movie_success_rate.csv")
```

In [74]:

```
df
```

Out[74]:

	Rank	Title	Genre	Description	Director	Actors	Year
0	1.0	Guardians of the Galaxy	Action,Adventure,Sci-Fi	A group of intergalactic criminals are forced ...	James Gunn	Chris Pratt, Vin Diesel, Bradley Cooper, Zoe S...	2014.0
1	2.0	Prometheus	Adventure,Mystery,Sci-Fi	Following clues to the origin of mankind, a te...	Ridley Scott	Noomi Rapace, Logan Marshall-Green, Michael Fa...	2012.0
2	3.0	Split	Horror,Thriller	Three girls are kidnapped by a man with a diag...	M. Night Shyamalan	James McAvoy, Anya Taylor-Joy, Haley Lu Richar...	2016.0
3	4.0	Sing	Animation,Comedy,Family	In a city of humanoid animals, a hustling thea...	Christophe Lourdelet	Matthew McConaughey,Reese Witherspoon, Seth Ma...	2016.0
4	5.0	Suicide Squad	Action,Adventure,Fantasy	A secret government agency recruits some of th...	David Ayer	Will Smith, Jared Leto, Margot Robbie, Viola D...	2016.0
...	...	...	...	...	...	...	...
834	995.0	Project X	Comedy	3 high school seniors throw a	Nima Nourizadeh	Thomas Mann, Oliver Cooper, Jonathan Daniel Br...	2012.0

	Rank	Title	Genre	Description	Director	Actors	Year
				birthday party t...			
835	997.0	Hostel: Part II	Horror	Three American college students studying abroa...	Eli Roth	Lauren German, Heather Matarazzo, Bijou Philli...	2007.0
836	998.0	Step Up 2: The Streets	Drama,Music,Romance	Romantic sparks occur between two dance studen...	Jon M. Chu	Robert Hoffman, Briana Evigan, Cassie Ventura,...	2008.0
837	1000.0	Nine Lives	Comedy,Family,Fantasy	A stuffy businessman finds himself trapped ins...	Barry Sonnenfeld	Kevin Spacey, Jennifer Garner, Robbie Amell,Ch...	2016.0
838	NaN	NaN	NaN	NaN	NaN	NaN	NaN

839 rows × 33 columns

In [76]:

```
print(df.columns)
```

```
Index(['Rank', 'Title', 'Genre', 'Description', 'Director', 'Actors', 'Year',
       'Runtime (Minutes)', 'Rating', 'Votes', 'Revenue (Millions)',
       'Metascore', 'Action', 'Adventure', 'Aniimation', 'Biography', 'Comedy',
       'Crime', 'Drama', 'Family', 'Fantasy', 'History', 'Horror', 'Music',
       'Musical', 'Mystery', 'Romance', 'Sci-Fi', 'Sport', 'Thriller', 'War',
       'Western', 'Success'],
      dtype='object')
```

In [78]:

```
df.isnull()
```

Out[78]:

	Rank	Title	Genre	Description	Director	Actors	Year	Runtime (Minutes)	Rating	Votes	...	Music	Mu
0	False	False	False	False	False	False	False	False	False	False	...	False	F
1	False	False	False	False	False	False	False	False	False	False	...	False	F
2	False	False	False	False	False	False	False	False	False	False	...	False	F
3	False	False	False	False	False	False	False	False	False	False	...	False	F
4	False	False	False	False	False	False	False	False	False	False	...	False	F
...	...	...	...	...	...	...	...	...	...	...	...	...	
834	False	False	False	False	False	False	False	False	False	False	...	False	F
835	False	False	False	False	False	False	False	False	False	False	...	False	F
836	False	False	False	False	False	False	False	False	False	False	...	False	F
837	False	False	False	False	False	False	False	False	False	False	...	False	F

	Rank	Title	Genre	Description	Director	Actors	Year	Runtime (Minutes)	Rating	Votes	...	Music	Mu
838	True	True	True	True	True	True	True	True	False	False	...	True	

839 rows × 33 columns

In [80]:

```
df.dropna()
```

Out[80]:

	Rank	Title	Genre	Description	Director	Actors	Year
0	1.0	Guardians of the Galaxy	Action,Adventure,Sci-Fi	A group of intergalactic criminals are forced ...	James Gunn	Chris Pratt, Vin Diesel, Bradley Cooper, Zoe S...	2014.0
1	2.0	Prometheus	Adventure,Mystery,Sci-Fi	Following clues to the origin of mankind, a te...	Ridley Scott	Noomi Rapace, Logan Marshall-Green, Michael Fa...	2012.0
2	3.0	Split	Horror,Thriller	Three girls are kidnapped by a man with a diag...	M. Night Shyamalan	James McAvoy, Anya Taylor-Joy, Haley Lu Richar...	2016.0
3	4.0	Sing	Animation,Comedy,Family	In a city of humanoid animals, a hustling thea...	Christophe Lourdelet	Matthew McConaughey,Reese Witherspoon, Seth Ma...	2016.0
4	5.0	Suicide Squad	Action,Adventure,Fantasy	A secret government agency recruits some of th...	David Ayer	Will Smith, Jared Leto, Margot Robbie, Viola D...	2016.0
...	...	...	...	...	...	...	...
833	994.0	Resident Evil: Afterlife	Action,Adventure,Horror	While still out to destroy the evil Umbrella C...	Paul W.S. Anderson	Milla Jovovich, Ali Larter, Wentworth Miller,K...	2010.0
834	995.0	Project X	Comedy	3 high school seniors throw a birthday party t...	Nima Nourizadeh	Thomas Mann, Oliver Cooper, Jonathan Daniel Br...	2012.0
835	997.0	Hostel: Part II	Horror	Three American college students studying abroa...	Eli Roth	Lauren German, Heather Matarazzo, Bijou Philli...	2007.0
836	998.0	Step Up 2: The Streets	Drama,Music,Romance	Romantic sparks occur between two	Jon M. Chu	Robert Hoffman, Briana Evigan, Cassie Ventura,...	2008.0

	Rank	Title	Genre	Description	Director	Actors	Year
				dance studen...			
837	1000.0	Nine Lives	Comedy,Family,Fantasy	A stuffy businessman finds himself trapped ins...	Barry Sonnenfeld	Kevin Spacey, Jennifer Garner, Robbie Amell,Ch...	2016.0

838 rows × 33 columns

In [82]:
```
df.rename(columns={'Aniimation':'Animation'},inplace=True)
```

In [84]:
```
string_columns=['Title','Genre','Description','Director','Actors']
df[string_columns]=df[string_columns].apply(lambda x:x.str.strip())
```

In [86]:
```
df=df[df['Year'].apply(lambda x:str(x).isdigit())]
df=df[df['Rank'].apply(lambda x:str(x).isdigit())]
```

In [88]:
```
df=df[~df.isin([np.inf,-np.inf]).any(axis=1)]
```

In [96]:
```
df['Year']=df['Year'].astype(int)
df['Rank']=df['Rank'].astype(int)
```

In [100]:
```
df.reset_index(drop=True,inplace=True)
```

In [102]:
```
print("Data cleaned")
print(df.info())
print(df.head(2))
```

```
Data cleaned
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 838 entries, 0 to 837
Data columns (total 33 columns):
 #   Column              Non-Null Count  Dtype
---  ------              --------------  -----
 0   Rank                838 non-null    int32
 1   Title               838 non-null    object
 2   Genre               838 non-null    object
 3   Description         838 non-null    object
 4   Director            838 non-null    object
 5   Actors              838 non-null    object
 6   Year                838 non-null    int32
 7   Runtime (Minutes)   838 non-null    float64
 8   Rating              838 non-null    float64
 9   Votes               838 non-null    float64
 10  Revenue (Millions)  838 non-null    float64
 11  Metascore           838 non-null    float64
 12  Action              838 non-null    float64
 13  Adventure           838 non-null    float64
```

```
 14  Aniimation          838 non-null    float64
 15  Biography           838 non-null    float64
 16  Comedy              838 non-null    float64
 17  Crime               838 non-null    float64
 18  Drama               838 non-null    float64
 19  Family              838 non-null    float64
 20  Fantasy             838 non-null    float64
 21  History             838 non-null    float64
 22  Horror              838 non-null    float64
 23  Music               838 non-null    float64
 24  Musical             838 non-null    float64
 25  Mystery             838 non-null    float64
 26  Romance             838 non-null    float64
 27  Sci-Fi              838 non-null    float64
 28  Sport               838 non-null    float64
 29  Thriller            838 non-null    float64
 30  War                 838 non-null    float64
 31  Western             838 non-null    float64
 32  Success             838 non-null    float64
dtypes: float64(26), int32(2), object(5)
memory usage: 209.6+ KB
None
   Rank                    Title                     Genre  \
0     1  Guardians of the Galaxy   Action,Adventure,Sci-Fi
1     2               Prometheus  Adventure,Mystery,Sci-Fi

                                         Description      Director  \
0  A group of intergalactic criminals are forced ...    James Gunn
1  Following clues to the origin of mankind, a te...  Ridley Scott

                                              Actors  Year  Runtime (Minutes)  \
0  Chris Pratt, Vin Diesel, Bradley Cooper, Zoe S...  2014              121.0
1  Noomi Rapace, Logan Marshall-Green, Michael Fa...  2012              124.0

   Rating     Votes  ...  Music  Musical  Mystery  Romance  Sci-Fi  Sport  \
0     8.1  757074.0  ...    0.0      0.0      0.0      0.0     1.0    0.0
1     7.0  485820.0  ...    0.0      0.0      1.0      0.0     1.0    0.0

   Thriller  War  Western  Success
0       0.0  0.0      0.0      1.0
1       0.0  0.0      0.0      1.0

[2 rows x 33 columns]
```

In [110]:

```
sid=SentimentIntensityAnalyzer()
```

In [116]:

```
df['sentiment'] = df['Description'].apply(lambda x: sid.polarity_scores(x)['compound'])
```

In [122]:

```
drop_columns = ['Title', 'Genre', 'Description', 'Director', 'Actors']
df_model = df.drop(columns=drop_columns)
```

In [131]:

```
x=df_model.drop(columns=['Success'])
y=df_model['Success']
```

In [133]:

```
x_train,x_test,y_train,y_test=train_test_split(x,y,test_size=0.2,random_state=42)
```

In [135]:

```
lr=LinearRegression()
lr.fit(x_train,y_train)
```

Out[135]:

LinearRegression

LinearRegression()

In [137]:

```
y_pred=lr.predict(x_test)
```

In [139]:

```
r2=r2_score(y_test,y_pred)
mae=mean_absolute_error(y_test,y_pred)
```

In [141]:

```
print("R2_Score:",r2)
print("Mean Absolute Error:",mae)
```

```
R2_Score: 0.4686220326933619
Mean Absolute Error: 0.2189490388311545
```

In [143]:

```
coeff_df=pd.DataFrame({
    'Feature':x.columns,
    'Coefficient':lr.coef_})
```

In [145]:

```
results_df=pd.DataFrame({
    'Actual':y_test.values,
    'predicted':y_pred})
```

In [151]:

```
with pd.ExcelWriter("regression_output.xlsx") as writer:
    coeff_df.to_excel(writer,sheet_name='Coefficients',index=False)
    results_df.to_excel(writer,sheet_name='Predictions',index=False)
    pd.DataFrame({'R2 Score': [r2], 'MAE': [mae]}).to_excel(writer, sheet_name='Metrics'
```

In [153]:

```
print("regression output saved to'regression_output.xlsx'")
```

```
regression output saved to'regression_output.xlsx'
```

In [ ]: