Movie Studio Analysis

Business Problem

Your company now sees all the big companies creating original video content and they want to get in on the fun. They have decided to create a new movie

studio, but they don’t know anything about creating movies. You are charged with exploring what types of films are currently doing the best at the box

office. You must then translate those findings into actionable insights that the head of your company's new movie studio can use to help decide what type

of films to create.

1. Business Understanding

Key Stakeholders: Company head

Objectives

Main Objective: To Explore the type of films that are currently doing the best at the box office to help create a new movie studio.

Specific Objectives:

1. To identify the most preffered films.
2. To analyze trends in the film industry.
3. To identify which movie genre performs best in terms of popularity and revenue.
4. To identify which genre has the highest ratings.

2. Data Understanding

Below are movie datasets in different formats collected from various locations:

- CSV(comma-separated values) file "tmdb.movies.csv.gz"
- TSV(tab-separated values) file "rt.movie_info.csv.gz".

3. Data Preparation

- Opening and inspecting the contents of csv file("tmdb.movies.csv.gz")
- Opening and inspecting the contents of tsv file("rt.movie_info.tsv.gz")
- Dealing with missing values
- Dealing with duplicates
- Performing EDA

Data Reading

```
# importing the necessary libraries
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
%matplotlib inline
```

In the cell below we;

load our datasets

```
tmdb_df = pd.read_csv("tmdb.movies.csv.gz")

rt_df = pd.read_csv("rt.movie_info.tsv.gz", sep ='\t')
```

Data Exploration

```
rt_df.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1560 entries, 0 to 1559
Data columns (total 12 columns):
 #   Column        Non-Null Count  Dtype 
---  ------        --------------  ----- 
 0   id            1560 non-null   int64 
 1   synopsis      1498 non-null   object
 2   rating        1557 non-null   object
 3   genre         1552 non-null   object
 4   director      1361 non-null   object
 5   writer        1111 non-null   object
 6   theater_date  1201 non-null   object
 7   dvd_date      1201 non-null   object
 8   currency      340 non-null    object
 9   box_office    340 non-null    object
 10  runtime       1530 non-null   object
 11  studio        494 non-null    object
```

```
dtypes: int64(1), object(11)
memory usage: 146.4+ KB

#genre column contents
rt_df["genre"]

0                      Action and Adventure|Classics|Drama
1                        Drama|Science Fiction and Fantasy
2                        Drama|Musical and Performing Arts
3                               Drama|Mystery and Suspense
4                                            Drama|Romance
                              ...
1555     Action and Adventure|Horror|Mystery and Suspense
1556                    Comedy|Science Fiction and Fantasy
1557    Classics|Comedy|Drama|Musical and Performing Arts
1558      Comedy|Drama|Kids and Family|Sports and Fitness
1559    Action and Adventure|Art House and Internation...
Name: genre, Length: 1560, dtype: object

# Exploring tmdb columns
tmdb_df.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 26517 entries, 0 to 26516
Data columns (total 10 columns):
 #   Column             Non-Null Count  Dtype
---  ------             --------------  -----
 0   Unnamed: 0         26517 non-null  int64
 1   genre_ids          26517 non-null  object
 2   id                 26517 non-null  int64
 3   original_language  26517 non-null  object
 4   original_title     26517 non-null  object
 5   popularity         26517 non-null  float64
 6   release_date       26517 non-null  object
 7   title              26517 non-null  object
 8   vote_average       26517 non-null  float64
 9   vote_count         26517 non-null  int64
dtypes: float64(2), int64(3), object(5)
memory usage: 2.0+ MB

# Merging the two datasets
movie_df = pd.merge(tmdb_df , rt_df, on="id")

# Previewing the merged datasets
movie_df.head()

   Unnamed: 0           genre_ids   id original_language
original_title  \
0           3     [16, 35, 10751]  862                en         Toy
Story
1          10     [16, 35, 10751]  863                en       Toy
```

```
Story 2
2          32   [28, 53, 878, 12]   95                en
Armageddon
3          43         [35, 10749]  239                en  Some Like It
Hot
4         117  [18, 10402, 10749]   27                en           9
Songs

   popularity release_date             title  vote_average  vote_count
...  \
0      28.005   1995-11-22         Toy Story           7.9       10174
...
1      22.698   1999-11-24       Toy Story 2           7.5        7553
...
2      15.799   1998-07-01        Armageddon           6.7        4267
...
3      14.200   1959-03-18  Some Like It Hot           8.2        1562
...
4      10.332   2004-09-09           9 Songs           4.9         170
...

  rating                                              genre  \
0  PG-13                                             Comedy
1      R  Action and Adventure|Art House and Internation...
2      R                           Drama|Sports and Fitness
3     PG                                      Comedy|Horror
4     NR                       Musical and Performing Arts

                  director          writer  theater_date      dvd_date
\
0  Anthony Russo|Joe Russo             NaN  Jul 13, 2006  Nov 21, 2006

1           Harmony Korine  Harmony Korine  Mar 22, 2013   Jul 9, 2013

2              Ben Younger     Ben Younger  Nov 18, 2016  Feb 14, 2017

3                      NaN             NaN           NaN           NaN

4                      NaN             NaN           NaN           NaN


  currency  box_office      runtime             studio
0        $  75,604,320  109 minutes  Universal Pictures
1        $  13,900,000   93 minutes           A24 Films
2        $   5,051,927  116 minutes     Open Road Films
3      NaN         NaN   80 minutes                 NaN
4      NaN         NaN          NaN                 NaN

[5 rows x 21 columns]
```

```python
# Previewing the merged dataset columns
movie_df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 32 entries, 0 to 31
Data columns (total 21 columns):
 #   Column             Non-Null Count  Dtype
---  ------             --------------  -----
 0   Unnamed: 0         32 non-null     int64
 1   genre_ids          32 non-null     object
 2   id                 32 non-null     int64
 3   original_language  32 non-null     object
 4   original_title     32 non-null     object
 5   popularity         32 non-null     float64
 6   release_date       32 non-null     object
 7   title              32 non-null     object
 8   vote_average       32 non-null     float64
 9   vote_count         32 non-null     int64
 10  synopsis           29 non-null     object
 11  rating             32 non-null     object
 12  genre              32 non-null     object
 13  director           29 non-null     object
 14  writer             21 non-null     object
 15  theater_date       24 non-null     object
 16  dvd_date           24 non-null     object
 17  currency           9 non-null      object
 18  box_office         9 non-null      object
 19  runtime            31 non-null     object
 20  studio             11 non-null     object
dtypes: float64(2), int64(3), object(16)
memory usage: 5.4+ KB
```

```python
# Checking the merged dataset dimensions
movie_df.shape
```

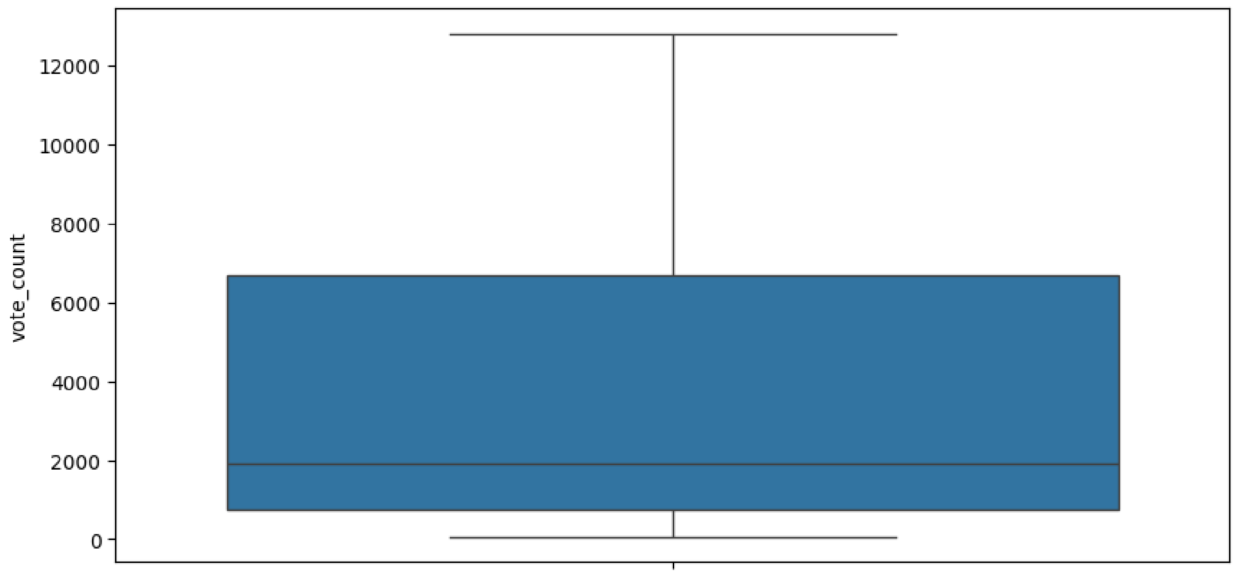
```
(32, 21)
```

Data Cleaning

Detecting outliers

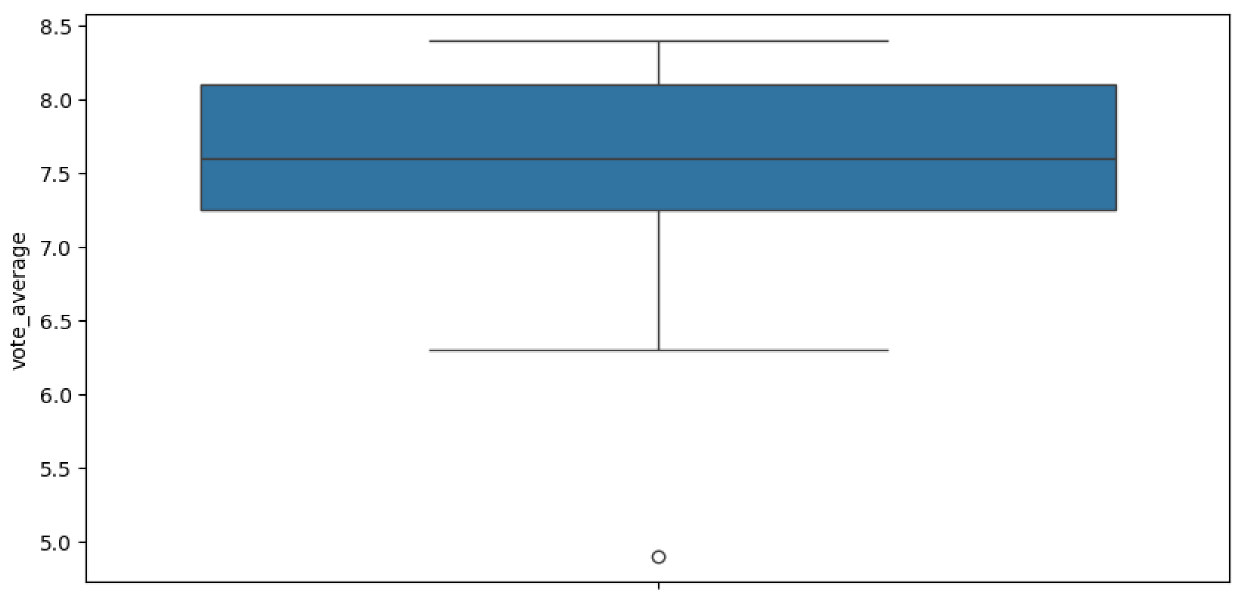
```python
# checking for outliers
plt.figure(figsize=(10,5))
sns.boxplot(movie_df['vote_count'])
```

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



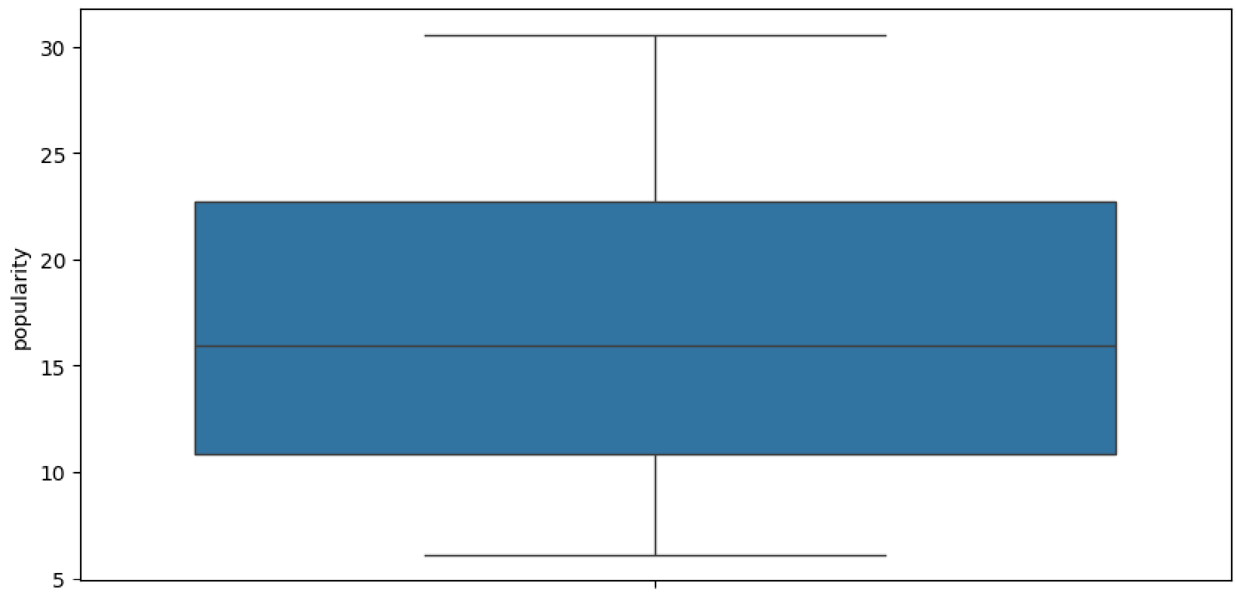

```
plt.figure(figsize=(10,5))
sns.boxplot(movie_df["vote_average"])
```

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




```
plt.figure(figsize=(10,5))
sns.boxplot(movie_df["popularity"])
```

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



There is almost no detected outliers.

Checking for duplicates

```
# Checking for duplicates.
movie_df.duplicated().sum()
```

```
np.int64(0)
```

```
movie_df
```

```
    Unnamed: 0            genre_ids    id original_language  \
0            3      [16, 35, 10751]   862                en
1           10      [16, 35, 10751]   863                en
2           32    [28, 53, 878, 12]    95                en
3           43          [35, 10749]   239                en
4          117  [18, 10402, 10749]    27                en
5          120                [878]   830                en
6         2470         [12, 28, 14]  1865                en
7         2473      [16, 35, 10751]   862                en
8         2474        [28, 12, 878]  1771                en
9         2477      [16, 35, 10751]   863                en
10        2485                 [18]   489                en
11        2494      [18, 36, 10752]   387                de
12        2500     [27, 28, 53, 80]   755                en
13        2510         [28, 35, 80]    90                en
14        2594          [18, 10402]   786                en
15        5186         [28, 12, 14]  1930                en
16        5192  [18, 9648, 53, 14]  1813                en
17        5201             [18, 80]   311                en
18       11047             [80, 53]   189                en
19       11109  [80, 18, 9648, 53]    93                en
```

```
20      11192     [18, 36, 10749]   887                 en
21      14222     [18, 36, 10752]   387                 de
22      14396     [28, 53, 10770]   839                 en
23      17395       [28, 53, 878]   280                 en
24      17932        [35, 18, 36]   986                 en
25      20639       [28, 53, 878]   280                 en
26      20745           [878, 18]   840                 en
27      24000         [35, 10749]   239                 en
28      24022                [18]   797                 sv
29      24186                [18]   221                 en
30      24211                [18]   614                 sv
31      24268            [14, 18]   490                 sv

                                 original_title  popularity
release_date  \
0                                     Toy Story      28.005   1995-11-
22
1                                   Toy Story 2      22.698   1999-11-
24
2                                    Armageddon      15.799   1998-07-
01
3                              Some Like It Hot      14.200   1959-03-
18
4                                       9 Songs      10.332   2004-09-
09
5                              Forbidden Planet      10.274   1956-03-
15
6   Pirates of the Caribbean: On Stranger Tides      30.579   2011-05-
20
7                                     Toy Story      28.005   1995-11-
22
8            Captain America: The First Avenger      25.808   2011-07-
22
9                                   Toy Story 2      22.698   1999-11-
24
10                            Good Will Hunting      18.013   1997-12-
05
11                                     Das Boot      16.554   1982-02-
10
12                           From Dusk Till Dawn      16.064   1996-01-
19
13                             Beverly Hills Cop      15.067   1984-11-
30
14                               Almost Famous      11.022   2000-09-
15
15                       The Amazing Spider-Man      24.391   2012-07-
04
16                         The Devil's Advocate      19.903   1997-10-
17
```

```
17                      Once Upon a Time in America      17.717   1984-06-
01
18                     Sin City: A Dame to Kill For      20.896   2014-08-
22
19                             Anatomy of a Murder       12.710   1959-07-
01
20                      The Best Years of Our Lives       9.647   1946-12-
25
21                                         Das Boot      16.554   1982-02-
10
22                                             Duel       8.661   2003-08-
12
23                       Terminator 2: Judgment Day      24.604   1991-07-
03
24                                         Falstaff       6.108   1967-03-
16
25                       Terminator 2: Judgment Day      24.604   1991-07-
03
26             Close Encounters of the Third Kind        13.044   1977-11-
16
27                                Some Like It Hot       14.200   1959-03-
18
28                                          Persona      13.342   1967-03-
06
29                          Rebel Without a Cause         9.752   2018-09-
23
30                                 Smultronstället        9.381   1957-12-
26
31                             Det sjunde inseglet        8.693   1958-10-
13

                                             title  vote_average
vote_count  \
0                                        Toy Story           7.9
10174
1                                      Toy Story 2           7.5
7553
2                                       Armageddon           6.7
4267
3                                 Some Like It Hot           8.2
1562
4                                          9 Songs           4.9
170
5                                 Forbidden Planet           7.3
388
6   Pirates of the Caribbean: On Stranger Tides              6.4
8571
7                                        Toy Story           7.9
10174
```

```
8              Captain America: The First Avenger              6.9
12810
9                                     Toy Story 2              7.5
7553
10                              Good Will Hunting              8.1
5764
11                                       Das Boot              8.1
981
12                            From Dusk Till Dawn              7.0
3015
13                              Beverly Hills Cop              7.1
1827
14                                 Almost Famous               7.5
1339
15                         The Amazing Spider-Man              6.5
10411
16                          The Devil's Advocate               7.3
2622
17                     Once Upon a Time in America             8.4
2243
18                    Sin City: A Dame to Kill For             6.3
2210
19                           Anatomy of a Murder               7.9
359
20                     The Best Years of Our Lives             7.8
243
21                                       Das Boot              8.1
981
22                                           Duel              7.4
742
23                       Terminator 2: Judgment Day            7.9
6682
24                             Chimes at Midnight              7.4
65
25                       Terminator 2: Judgment Day            7.9
6682
26             Close Encounters of the Third Kind              7.3
2005
27                               Some Like It Hot              8.2
1562
28                                        Persona              8.3
726
29                          Rebel Without a Cause              7.7
740
30                              Wild Strawberries              8.1
595
31                                The Seventh Seal             8.2
1163
```

```
    ... rating                                              genre  \
0   ...  PG-13                                             Comedy   
1   ...      R  Action and Adventure|Art House and Internation...   
2   ...      R                          Drama|Sports and Fitness   
3   ...     PG                                      Comedy|Horror   
4   ...     NR                       Musical and Performing Arts   
5   ...      R  Art House and International|Comedy|Drama|Roman...   
6   ...      R                                              Drama   
7   ...  PG-13                                             Comedy   
8   ...     NR                         Action and Adventure|Drama   
9   ...      R  Action and Adventure|Art House and Internation...   
10  ...      R                                             Comedy   
11  ...     NR     Action and Adventure|Classics|Western|Romance   
12  ...  PG-13   Action and Adventure|Science Fiction and Fantasy   
13  ...     NR          Drama|Musical and Performing Arts|Romance   
14  ...     NR        Action and Adventure|Drama|Special Interest   
15  ...     PG  Action and Adventure|Art House and Internation...   
16  ...      G          Drama|Horror|Science Fiction and Fantasy   
17  ...  PG-13                 Art House and International|Drama   
18  ...     NR     Drama|Horror|Mystery and Suspense|Television   
19  ...      R                     Classics|Comedy|Drama|Romance   
20  ...      R                              Comedy|Drama|Romance   
21  ...     NR     Action and Adventure|Classics|Western|Romance   
22  ...     PG                                      Drama|Romance   
23  ...     NR                        Art House and International   
24  ...      R                                              Drama   
25  ...     NR                        Art House and International   
26  ...     NR            Classics|Drama|Mystery and Suspense   
27  ...     PG                                      Comedy|Horror   
28  ...     PG                                              Drama   
29  ...     PG                                     Classics|Drama   
30  ...     PG                             Classics|Drama|Western   
31  ...     NR   Action and Adventure|Art House and International   

                     director  \
0         Anthony Russo|Joe Russo   
1                  Harmony Korine   
2                     Ben Younger   
3                             NaN   
4                             NaN   
5                         Ang Lee   
6                    Craig Brewer   
7         Anthony Russo|Joe Russo   
8                      Fritz Lang   
9                  Harmony Korine   
10                      Tony Bill   
11                      Tom Gries   
12                 Richard Donner   
13                 Herbert Wilcox   
```

14	Jarrett Lee Conaway
15	Guy Hamilton
16	Kinji Fukasaku
17	Paolo Taviani|Vittorio Taviani
18	Daniel Sackheim
19	Ernst Lubitsch
20	Charles Shyer
21	Tom Gries
22	Arthur Hiller
23	Lo Po-Shan
24	Martin Scorsese
25	Lo Po-Shan
26	Mervyn Le Roy
27	NaN
28	Charles Burnett
29	Richard Brooks
30	Mark Rydell
31	Rohit Shetty

	writer	theater_date \
0	NaN	Jul 13, 2006
1	Harmony Korine	Mar 22, 2013
2	Ben Younger	Nov 18, 2016
3	NaN	NaN
4	NaN	NaN
5	Ang Lee|James Schamus|Neil Peng	Aug 4, 1993
6	Craig Brewer	Mar 2, 2007
7	NaN	Jul 13, 2006
8	Jan Lustig|Margaret Fitts	Jan 1, 1955
9	Harmony Korine	Mar 22, 2013
10	Mitch Markowitz	Apr 11, 1990
11	Tom Gries	Apr 10, 1968
12	Jeff Maguire|George Nolfi	Nov 26, 2003
13	Ken Englund	NaN
14	NaN	NaN
15	Evan Jones	Jan 1, 1966
16	Charles Sinclair|Tom Rowe|William Finger	Jan 1, 1969
17	Paolo Taviani|Vittorio Taviani|Sandro Petraglia	Jan 1, 1993
18	Anthony Spinner	Sep 29, 1996
19	Charles Brackett|Billy Wilder|Walter Reisch|Me...	Nov 3, 1939
20	Charles Shyer|Elaine Pope	Nov 5, 2004
21	Tom Gries	Apr 10, 1968
22	Erich Segal	Dec 16, 1970
23	NaN	NaN
24	Paul Schrader	Aug 12, 1988
25	NaN	NaN
26	Sheridan Gibney|Brown Holmes	Nov 19, 1932
27	NaN	NaN
28	Charles Burnett	Jan 1, 1990

29	NaN	Jun 1, 1957
30	NaN	Jan 13, 1972
31	NaN	NaN

	dvd_date	currency	box_office	runtime	studio
0	Nov 21, 2006	$	75,604,320	109 minutes	Universal Pictures
1	Jul 9, 2013	$	13,900,000	93 minutes	A24 Films
2	Feb 14, 2017	$	5,051,927	116 minutes	Open Road Films
3	NaN	NaN	NaN	80 minutes	NaN
4	NaN	NaN	NaN	NaN	NaN
5	Jun 15, 2004	NaN	NaN	111 minutes	NaN
6	Jun 26, 2007	$	9,262,318	115 minutes	Paramount Vantage
7	Nov 21, 2006	$	75,604,320	109 minutes	Universal Pictures
8	Jan 22, 1992	NaN	NaN	89 minutes	NaN
9	Jul 9, 2013	$	13,900,000	93 minutes	A24 Films
10	Jul 6, 2004	NaN	NaN	91 minutes	NaN
11	Jun 4, 2002	NaN	NaN	109 minutes	NaN
12	Apr 13, 2004	$	19,375,474	116 minutes	Paramount Pictures
13	NaN	NaN	NaN	96 minutes	NaN
14	NaN	NaN	NaN	23 minutes	NaN
15	Aug 14, 2001	NaN	NaN	102 minutes	NaN
16	Sep 25, 1991	NaN	NaN	90 minutes	NaN
17	Apr 1, 2008	NaN	NaN	120 minutes	NaN
18	Feb 4, 2003	NaN	NaN	94 minutes	NaN
19	Sep 5, 2005	NaN	NaN	110 minutes	NaN
20	Mar 15, 2005	$	13,351,235	105 minutes	Paramount Pictures
21	Jun 4, 2002	NaN	NaN	109 minutes	NaN
22	Apr 24, 2001	NaN	NaN	100 minutes	Paramount Pictures

```
23           NaN       NaN         NaN   89 minutes                  NaN

24  Apr 25, 2000       NaN         NaN  164 minutes  Universal Pictures

25           NaN       NaN         NaN   89 minutes                  NaN

26  May 10, 2005       NaN         NaN   90 minutes                  NaN

27           NaN       NaN         NaN   80 minutes                  NaN

28  Jun 13, 1991       NaN         NaN  102 minutes                  NaN

29  Dec 13, 2011       NaN         NaN  147 minutes                  NaN

30   Oct 6, 1998       NaN         NaN  128 minutes                  NaN

31           NaN         $   1,231,550  145 minutes  Eros Entertainment


[32 rows x 21 columns]
```

There are duplicates in the dataframe where some rows are the same.

```
rows_to_drop = [7, 9, 11, 23, 27]
movie_df = movie_df.drop(movie_df.index[rows_to_drop])


movie_df
```

```
    Unnamed: 0           genre_ids    id original_language  \
0            3     [16, 35, 10751]   862                en
1           10     [16, 35, 10751]   863                en
2           32   [28, 53, 878, 12]    95                en
3           43         [35, 10749]   239                en
4          117  [18, 10402, 10749]    27                en
5          120               [878]   830                en
6         2470        [12, 28, 14]  1865                en
8         2474       [28, 12, 878]  1771                en
10        2485                [18]   489                en
12        2500    [27, 28, 53, 80]   755                en
13        2510        [28, 35, 80]    90                en
14        2594         [18, 10402]   786                en
15        5186        [28, 12, 14]  1930                en
16        5192  [18, 9648, 53, 14]  1813                en
17        5201            [18, 80]   311                en
18       11047            [80, 53]   189                en
19       11109  [80, 18, 9648, 53]    93                en
20       11192     [18, 36, 10749]   887                en
21       14222     [18, 36, 10752]   387                de
22       14396     [28, 53, 10770]   839                en
```

```
24       17932        [35, 18, 36]   986                en
25       20639       [28, 53, 878]   280                en
26       20745           [878, 18]   840                en
28       24022                [18]   797                sv
29       24186                [18]   221                en
30       24211                [18]   614                sv
31       24268            [14, 18]   490                sv

                                original_title  popularity  
release_date  \
0                                    Toy Story      28.005   1995-11-
22
1                                  Toy Story 2      22.698   1999-11-
24
2                                   Armageddon      15.799   1998-07-
01
3                             Some Like It Hot      14.200   1959-03-
18
4                                      9 Songs      10.332   2004-09-
09
5                             Forbidden Planet      10.274   1956-03-
15
6   Pirates of the Caribbean: On Stranger Tides      30.579   2011-05-
20
8            Captain America: The First Avenger      25.808   2011-07-
22
10                           Good Will Hunting      18.013   1997-12-
05
12                         From Dusk Till Dawn      16.064   1996-01-
19
13                           Beverly Hills Cop      15.067   1984-11-
30
14                              Almost Famous      11.022   2000-09-
15
15                        The Amazing Spider-Man      24.391   2012-07-
04
16                         The Devil's Advocate      19.903   1997-10-
17
17                  Once Upon a Time in America      17.717   1984-06-
01
18                    Sin City: A Dame to Kill For      20.896   2014-08-
22
19                           Anatomy of a Murder      12.710   1959-07-
01
20                  The Best Years of Our Lives       9.647   1946-12-
25
21                                    Das Boot      16.554   1982-02-
10
22                                        Duel       8.661   2003-08-
```

```
12
24                                    Falstaff       6.108   1967-03-
16
25                  Terminator 2: Judgment Day      24.604   1991-07-
03
26          Close Encounters of the Third Kind      13.044   1977-11-
16
28                                     Persona      13.342   1967-03-
06
29                     Rebel Without a Cause        9.752   2018-09-
23
30                            Smultronstället       9.381   1957-12-
26
31                       Det sjunde inseglet        8.693   1958-10-
13

                                         title  vote_average
vote_count  \
0                                    Toy Story           7.9
10174
1                                  Toy Story 2           7.5
7553
2                                   Armageddon           6.7
4267
3                             Some Like It Hot           8.2
1562
4                                      9 Songs           4.9
170
5                             Forbidden Planet           7.3
388
6   Pirates of the Caribbean: On Stranger Tides           6.4
8571
8           Captain America: The First Avenger           6.9
12810
10                           Good Will Hunting           8.1
5764
12                         From Dusk Till Dawn           7.0
3015
13                           Beverly Hills Cop           7.1
1827
14                              Almost Famous           7.5
1339
15                      The Amazing Spider-Man           6.5
10411
16                        The Devil's Advocate           7.3
2622
17                 Once Upon a Time in America           8.4
2243
18                  Sin City: A Dame to Kill For           6.3
```

```
2210
19                   Anatomy of a Murder           7.9
359
20           The Best Years of Our Lives           7.8
243
21                              Das Boot           8.1
981
22                                  Duel           7.4
742
24                   Chimes at Midnight           7.4
65
25            Terminator 2: Judgment Day           7.9
6682
26     Close Encounters of the Third Kind          7.3
2005
28                               Persona           8.3
726
29                Rebel Without a Cause            7.7
740
30                     Wild Strawberries           8.1
595
31                       The Seventh Seal          8.2
1163

    ... rating                                               genre  \
0   ...  PG-13                                              Comedy
1   ...      R  Action and Adventure|Art House and Internation...
2   ...      R                          Drama|Sports and Fitness
3   ...     PG                                      Comedy|Horror
4   ...     NR                        Musical and Performing Arts
5   ...      R  Art House and International|Comedy|Drama|Roman...
6   ...      R                                               Drama
8   ...     NR                          Action and Adventure|Drama
10  ...      R                                              Comedy
12  ...  PG-13   Action and Adventure|Science Fiction and Fantasy
13  ...     NR          Drama|Musical and Performing Arts|Romance
14  ...     NR        Action and Adventure|Drama|Special Interest
15  ...     PG  Action and Adventure|Art House and Internation...
16  ...      G           Drama|Horror|Science Fiction and Fantasy
17  ...  PG-13                  Art House and International|Drama
18  ...     NR       Drama|Horror|Mystery and Suspense|Television
19  ...      R                      Classics|Comedy|Drama|Romance
20  ...      R                               Comedy|Drama|Romance
21  ...     NR     Action and Adventure|Classics|Western|Romance
22  ...     PG                                      Drama|Romance
24  ...      R                                               Drama
25  ...     NR                        Art House and International
26  ...     NR                Classics|Drama|Mystery and Suspense
28  ...     PG                                               Drama
```

```
29  ...     PG                                        Classics|Drama
30  ...     PG                                Classics|Drama|Western
31  ...     NR   Action and Adventure|Art House and International

                              director  \
0          Anthony Russo|Joe Russo
1                   Harmony Korine
2                      Ben Younger
3                              NaN
4                              NaN
5                          Ang Lee
6                     Craig Brewer
8                       Fritz Lang
10                       Tony Bill
12                  Richard Donner
13                  Herbert Wilcox
14             Jarrett Lee Conaway
15                    Guy Hamilton
16                  Kinji Fukasaku
17  Paolo Taviani|Vittorio Taviani
18                 Daniel Sackheim
19                  Ernst Lubitsch
20                   Charles Shyer
21                       Tom Gries
22                   Arthur Hiller
24                 Martin Scorsese
25                     Lo Po-Shan
26                   Mervyn Le Roy
28                 Charles Burnett
29                  Richard Brooks
30                     Mark Rydell
31                    Rohit Shetty

                                               writer  theater_date  \
0                                                 NaN  Jul 13, 2006
1                                      Harmony Korine  Mar 22, 2013
2                                         Ben Younger  Nov 18, 2016
3                                                 NaN           NaN
4                                                 NaN           NaN
5                     Ang Lee|James Schamus|Neil Peng   Aug 4, 1993
6                                        Craig Brewer   Mar 2, 2007
8                           Jan Lustig|Margaret Fitts   Jan 1, 1955
10                                    Mitch Markowitz  Apr 11, 1990
12                          Jeff Maguire|George Nolfi  Nov 26, 2003
13                                        Ken Englund           NaN
14                                                NaN           NaN
15                                         Evan Jones   Jan 1, 1966
16          Charles Sinclair|Tom Rowe|William Finger   Jan 1, 1969
17    Paolo Taviani|Vittorio Taviani|Sandro Petraglia   Jan 1, 1993
```

18	Anthony Spinner	Sep 29, 1996
19	Charles Brackett|Billy Wilder|Walter Reisch|Me...	Nov 3, 1939
20	Charles Shyer|Elaine Pope	Nov 5, 2004
21	Tom Gries	Apr 10, 1968
22	Erich Segal	Dec 16, 1970
24	Paul Schrader	Aug 12, 1988
25	NaN	NaN
26	Sheridan Gibney|Brown Holmes	Nov 19, 1932
28	Charles Burnett	Jan 1, 1990
29	NaN	Jun 1, 1957
30	NaN	Jan 13, 1972
31	NaN	NaN

	dvd_date	currency	box_office	runtime	studio
0	Nov 21, 2006	$	75,604,320	109 minutes	Universal Pictures
1	Jul 9, 2013	$	13,900,000	93 minutes	A24 Films
2	Feb 14, 2017	$	5,051,927	116 minutes	Open Road Films
3	NaN	NaN	NaN	80 minutes	NaN
4	NaN	NaN	NaN	NaN	NaN
5	Jun 15, 2004	NaN	NaN	111 minutes	NaN
6	Jun 26, 2007	$	9,262,318	115 minutes	Paramount Vantage
8	Jan 22, 1992	NaN	NaN	89 minutes	NaN
10	Jul 6, 2004	NaN	NaN	91 minutes	NaN
12	Apr 13, 2004	$	19,375,474	116 minutes	Paramount Pictures
13	NaN	NaN	NaN	96 minutes	NaN
14	NaN	NaN	NaN	23 minutes	NaN
15	Aug 14, 2001	NaN	NaN	102 minutes	NaN
16	Sep 25, 1991	NaN	NaN	90 minutes	NaN
17	Apr 1, 2008	NaN	NaN	120 minutes	NaN
18	Feb 4, 2003	NaN	NaN	94 minutes	NaN
19	Sep 5, 2005	NaN	NaN	110 minutes	NaN
20	Mar 15, 2005	$	13,351,235	105 minutes	Paramount Pictures

```
21    Jun 4, 2002        NaN           NaN  109 minutes                    NaN
22   Apr 24, 2001        NaN           NaN  100 minutes   Paramount Pictures
24   Apr 25, 2000        NaN           NaN  164 minutes   Universal Pictures
25            NaN        NaN           NaN   89 minutes                    NaN
26   May 10, 2005        NaN           NaN   90 minutes                    NaN
28   Jun 13, 1991        NaN           NaN  102 minutes                    NaN
29   Dec 13, 2011        NaN           NaN  147 minutes                    NaN
30    Oct 6, 1998        NaN           NaN  128 minutes                    NaN
31            NaN          $    1,231,550  145 minutes   Eros Entertainment

[27 rows x 21 columns]
```

Dealing with missing values

```
# Identifying missing values
movie_df.isna().sum()
```

```
Unnamed: 0            0
genre_ids             0
id                    0
original_language     0
original_title        0
popularity            0
release_date          0
title                 0
vote_average          0
vote_count            0
synopsis              2
rating                0
genre                 0
director              2
writer                8
theater_date          6
dvd_date              6
currency             20
box_office           20
runtime               1
studio               18
dtype: int64
```

There are several missing values in our dataframe and we will deal with missing values by either droping or imputing a string to fill the null values

```
# Droping columns that have more missing values that are not required for my analysis
movie_df = movie_df.drop(columns = ["currency", "theater_date", "dvd_date","writer"])

# Box office contains numeric values but it's identified as an object dtype
movie_df["box_office"] = movie_df["box_office"].fillna(0)
movie_df["box_office"]
```

```
0     75,604,320
1     13,900,000
2      5,051,927
3              0
4              0
5              0
6      9,262,318
8              0
10             0
12    19,375,474
13             0
14             0
15             0
16             0
17             0
18             0
19             0
20    13,351,235
21             0
22             0
24             0
25             0
26             0
28             0
29             0
30             0
31     1,231,550
Name: box_office, dtype: object
```

```
#imputing missing values in the runtime column with unknown
movie_df["runtime"] = movie_df["runtime"].fillna("unknown")
movie_df["runtime"]
```

```
0     109 minutes
1      93 minutes
2     116 minutes
3      80 minutes
4         unknown
```

```
5     111 minutes
6     115 minutes
8      89 minutes
10     91 minutes
12    116 minutes
13     96 minutes
14     23 minutes
15    102 minutes
16     90 minutes
17    120 minutes
18     94 minutes
19    110 minutes
20    105 minutes
21    109 minutes
22    100 minutes
24    164 minutes
25     89 minutes
26     90 minutes
28    102 minutes
29    147 minutes
30    128 minutes
31    145 minutes
Name: runtime, dtype: object
```

```
#imputing missing values in the studio column with missing
movie_df["studio"] = movie_df["studio"].fillna("missing")
movie_df["studio"]
```

```
0     Universal Pictures
1              A24 Films
2        Open Road Films
3                missing
4                missing
5                missing
6      Paramount Vantage
8                missing
10               missing
12    Paramount Pictures
13               missing
14               missing
15               missing
16               missing
17               missing
18               missing
19               missing
20    Paramount Pictures
21               missing
22    Paramount Pictures
24    Universal Pictures
25               missing
```

```
26                missing
28                missing
29                missing
30                missing
31    Eros Entertainment
Name: studio, dtype: object
```

```
# Droping the unnamed column
movie_df = movie_df.loc[:, ~movie_df.columns.str.contains('Unnamed:
0')]
movie_df.head()
```

```
            genre_ids   id original_language    original_title
popularity  \
0     [16, 35, 10751]  862                en         Toy Story
28.005
1     [16, 35, 10751]  863                en       Toy Story 2
22.698
2   [28, 53, 878, 12]   95                en        Armageddon
15.799
3         [35, 10749]  239                en  Some Like It Hot
14.200
4  [18, 10402, 10749]   27                en           9 Songs
10.332

  release_date             title  vote_average  vote_count  \
0   1995-11-22         Toy Story           7.9       10174
1   1999-11-24       Toy Story 2           7.5        7553
2   1998-07-01        Armageddon           6.7        4267
3   1959-03-18  Some Like It Hot           8.2        1562
4   2004-09-09           9 Songs           4.9         170

                                            synopsis rating  \
0  A man is being driven crazy by his shiftless b...  PG-13
1  Brit (Ashley Benson), Candy (Vanessa Hudgens),...      R
2  BLEED FOR THIS is the incredible true story of...      R
3  In this film, a woman (Teri Garr) begins to st...     PG
4                                                NaN     NR

                                                 genre
director  \
0                                               Comedy  Anthony Russo|
Joe Russo
1  Action and Adventure|Art House and Internation...           Harmony
Korine
2                          Drama|Sports and Fitness               Ben
Younger
3                                        Comedy|Horror
NaN
4                         Musical and Performing Arts
```

```
NaN
```

	box_office	runtime	studio
0	75,604,320	109 minutes	Universal Pictures
1	13,900,000	93 minutes	A24 Films
2	5,051,927	116 minutes	Open Road Films
3	0	80 minutes	missing
4	0	unknown	missing

3. Data Analysis

```
# Summary statistics of numeric columns
movie_df.describe()
```

	id	popularity	vote_average	vote_count
count	27.000000	27.000000	27.000000	27.000000
mean	716.629630	16.046815	7.411111	3304.703704
std	564.255476	6.614108	0.781189	3668.227562
min	27.000000	6.108000	4.900000	65.000000
25%	259.500000	10.303000	7.050000	733.000000
50%	755.000000	15.067000	7.500000	1827.000000
75%	862.500000	20.399500	8.000000	5015.500000
max	1930.000000	30.579000	8.400000	12810.000000

```
# Removing commas and converting to float
movie_df["box_office"] = movie_df["box_office"].str.replace(",", "")
movie_df["box_office"] = pd.to_numeric(movie_df["box_office"],
errors="coerce")
movie_df["box_office"]
```

```
0     75604320.0
1     13900000.0
2      5051927.0
3            NaN
4            NaN
5            NaN
6      9262318.0
8            NaN
10           NaN
12    19375474.0
13           NaN
14           NaN
15           NaN
16           NaN
17           NaN
18           NaN
19           NaN
20    13351235.0
```

```
21          NaN
22          NaN
24          NaN
25          NaN
26          NaN
28          NaN
29          NaN
30          NaN
31    1231550.0
Name: box_office, dtype: float64
```

```
# Dropping the NaN values in my box_office column
movie_df = movie_df.dropna(subset=["box_office"])
movie_df
```

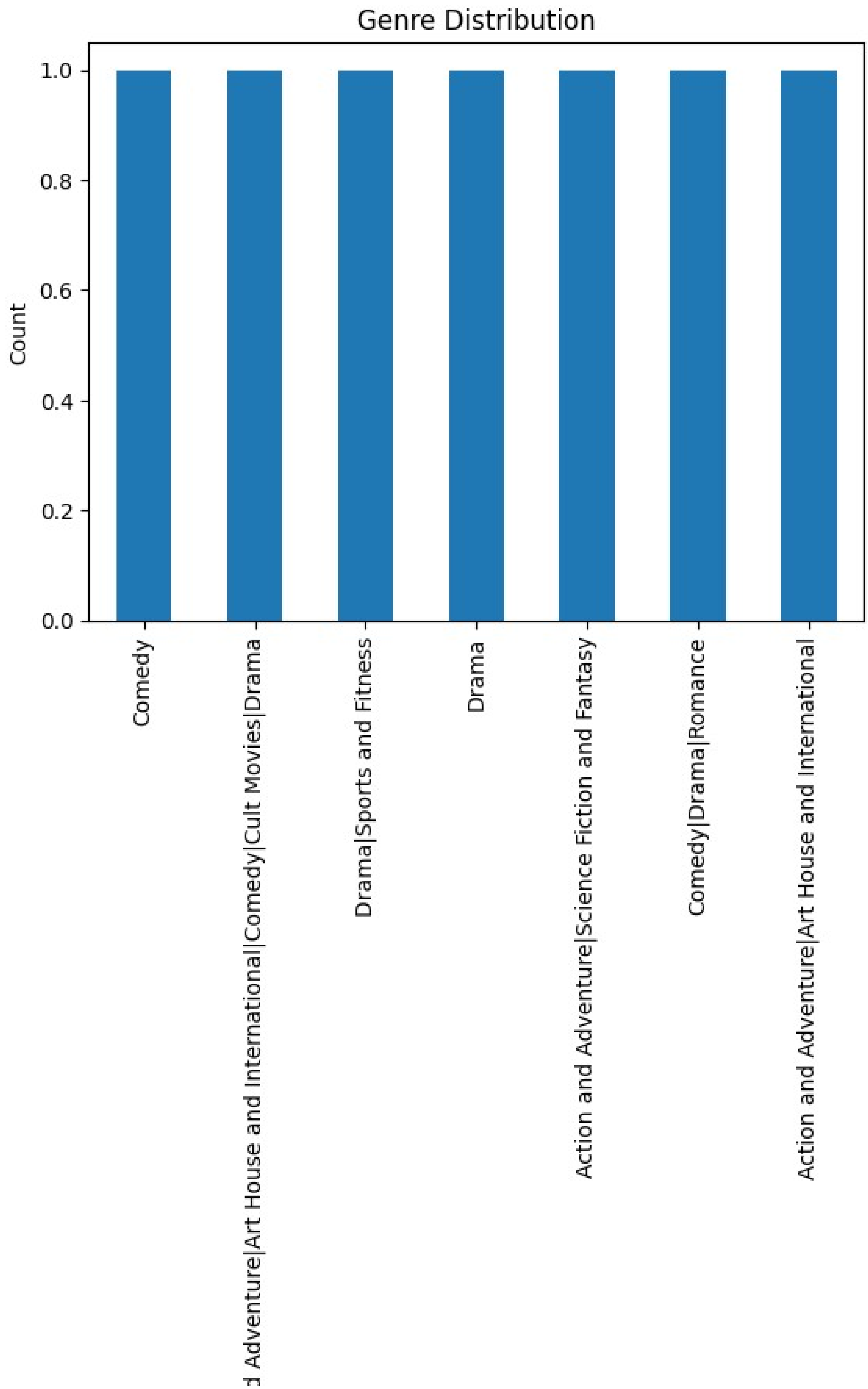
```
            genre_ids    id original_language  \
0     [16, 35, 10751]   862                en
1     [16, 35, 10751]   863                en
2   [28, 53, 878, 12]    95                en
6        [12, 28, 14]  1865                en
12   [27, 28, 53, 80]   755                en
20    [18, 36, 10749]   887                en
31           [14, 18]   490                sv

                                  original_title  popularity
release_date  \
0                                      Toy Story      28.005   1995-11-
22
1                                    Toy Story 2      22.698   1999-11-
24
2                                     Armageddon      15.799   1998-07-
01
6   Pirates of the Caribbean: On Stranger Tides      30.579   2011-05-
20
12                           From Dusk Till Dawn      16.064   1996-01-
19
20                   The Best Years of Our Lives       9.647   1946-12-
25
31                           Det sjunde inseglet       8.693   1958-10-
13

                                           title  vote_average
vote_count  \
0                                      Toy Story           7.9
10174
1                                    Toy Story 2           7.5
7553
2                                     Armageddon           6.7
4267
6   Pirates of the Caribbean: On Stranger Tides           6.4
```

```
8571
12                          From Dusk Till Dawn           7.0
3015
20                  The Best Years of Our Lives           7.8
243
31                             The Seventh Seal           8.2
1163

                                             synopsis rating  \
0   A man is being driven crazy by his shiftless b...  PG-13
1   Brit (Ashley Benson), Candy (Vanessa Hudgens),...      R
2   BLEED FOR THIS is the incredible true story of...      R
6   When a weathered, God-fearing ex-blues musicia...      R
12  Directing his first film since 1998's Lethal W...  PG-13
20  Alfie Elkins is a philosophical womanizer who ...      R
31  Following the 2011 super hit movie Singham, th...     NR

                                                genre  \
0                                              Comedy
1   Action and Adventure|Art House and Internation...
2                            Drama|Sports and Fitness
6                                               Drama
12   Action and Adventure|Science Fiction and Fantasy
20                               Comedy|Drama|Romance
31   Action and Adventure|Art House and International

                  director  box_office      runtime
studio
0   Anthony Russo|Joe Russo  75604320.0  109 minutes  Universal
Pictures
1             Harmony Korine  13900000.0   93 minutes          A24
Films
2                Ben Younger   5051927.0  116 minutes     Open Road
Films
6               Craig Brewer   9262318.0  115 minutes   Paramount
Vantage
12            Richard Donner  19375474.0  116 minutes  Paramount
Pictures
20             Charles Shyer  13351235.0  105 minutes  Paramount
Pictures
31              Rohit Shetty   1231550.0  145 minutes  Eros
Entertainment
```

Univariate Analysis

```
# getting value counts for the genre column and visualizing using a
histogram
movie_df['genre'].value_counts().plot(
    kind='bar',
```

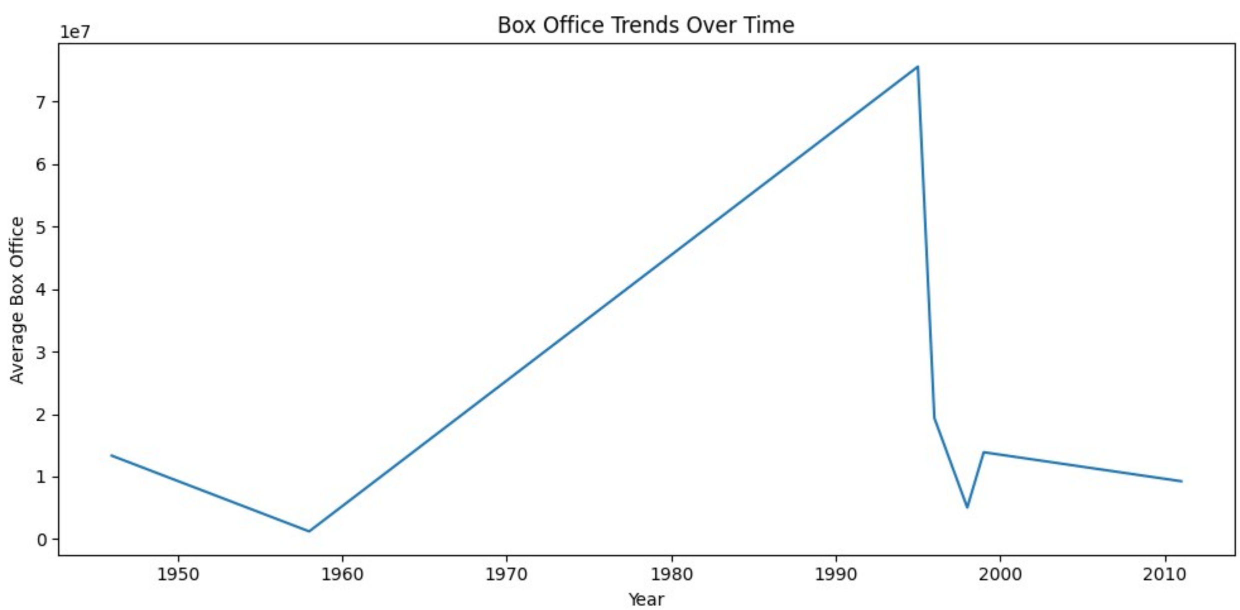
```
    rot=90,
)
#ploting
plt.title('Genre Distribution')
plt.xlabel('Genre')
plt.ylabel('Count')
plt.show()
```


Genre Distribution
Count
1.0
0.8
0.6
0.4
0.2
0.0
Comedy
d Adventure|Art House and International|Comedy|Cult Movies|Drama
Drama|Sports and Fitness
Drama
Action and Adventure|Science Fiction and Fantasy
Comedy|Drama|Romance
Action and Adventure|Art House and International

The genres are distribution equally in the box office.

Bivariate Analysis

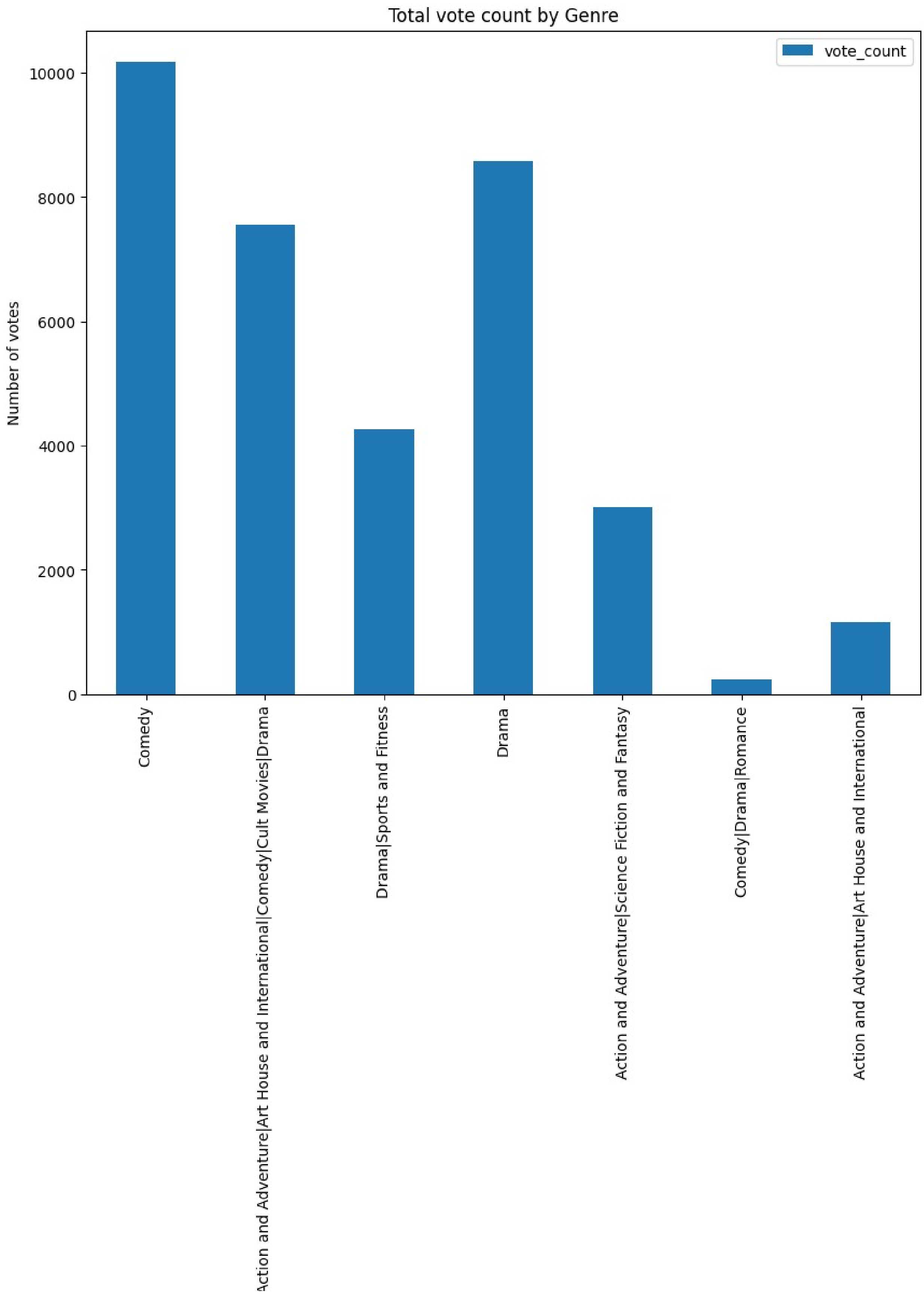
```
# Identifying box office trend overtime
#We will first parse release date to extract year
movie_df["release_date"] = pd.to_datetime(movie_df["release_date"],
errors="coerce")
movie_df["year"] = movie_df["release_date"].dt.year

yearly_box = movie_df.groupby("year")
["box_office"].mean().reset_index()

plt.figure(figsize=(10, 5))
sns.lineplot(data=yearly_box, x="year", y="box_office")
plt.title("Box Office Trends Over Time")
plt.xlabel("Year")
plt.ylabel("Average Box Office")
plt.tight_layout()
plt.show()
```



Looking at the most recent years, Average box_office has been a little bit sloppy.

```
#Visualizing the genre with the highest votes
ax = movie_df.plot.bar(
    x="genre",
    y="vote_count",
    figsize=(10, 8),
)
ax.set_title("Total vote count by Genre")
ax.set_xlabel("genres")
```

```
ax.set_ylabel("Number of votes")
plt.show()
```


Total vote count by Genre
vote_count
Number of votes
0
2000
4000
6000
8000
10000
Comedy
Action and Adventure|Art House and International|Comedy|Cult Movies|Drama
Drama|Sports and Fitness
Drama
Action and Adventure|Science Fiction and Fantasy
Comedy|Drama|Romance
Action and Adventure|Art House and International

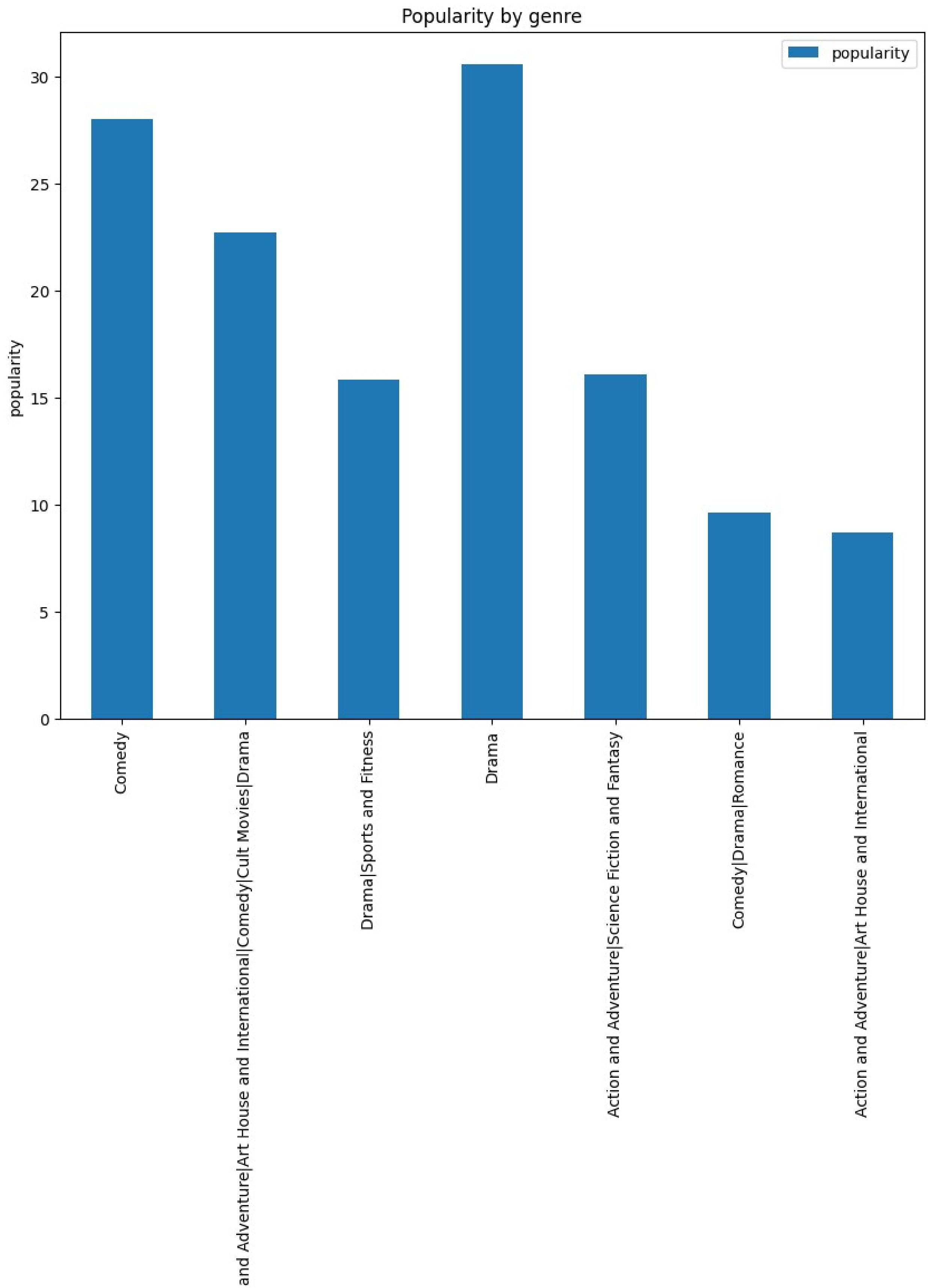
Comedy movie genre has the highest number of votes

Now let's take a look at the most popular genre by visualizing in terms of the popularity column

```
# A glimpse into the popularity data
movie_df["popularity"].head()
```

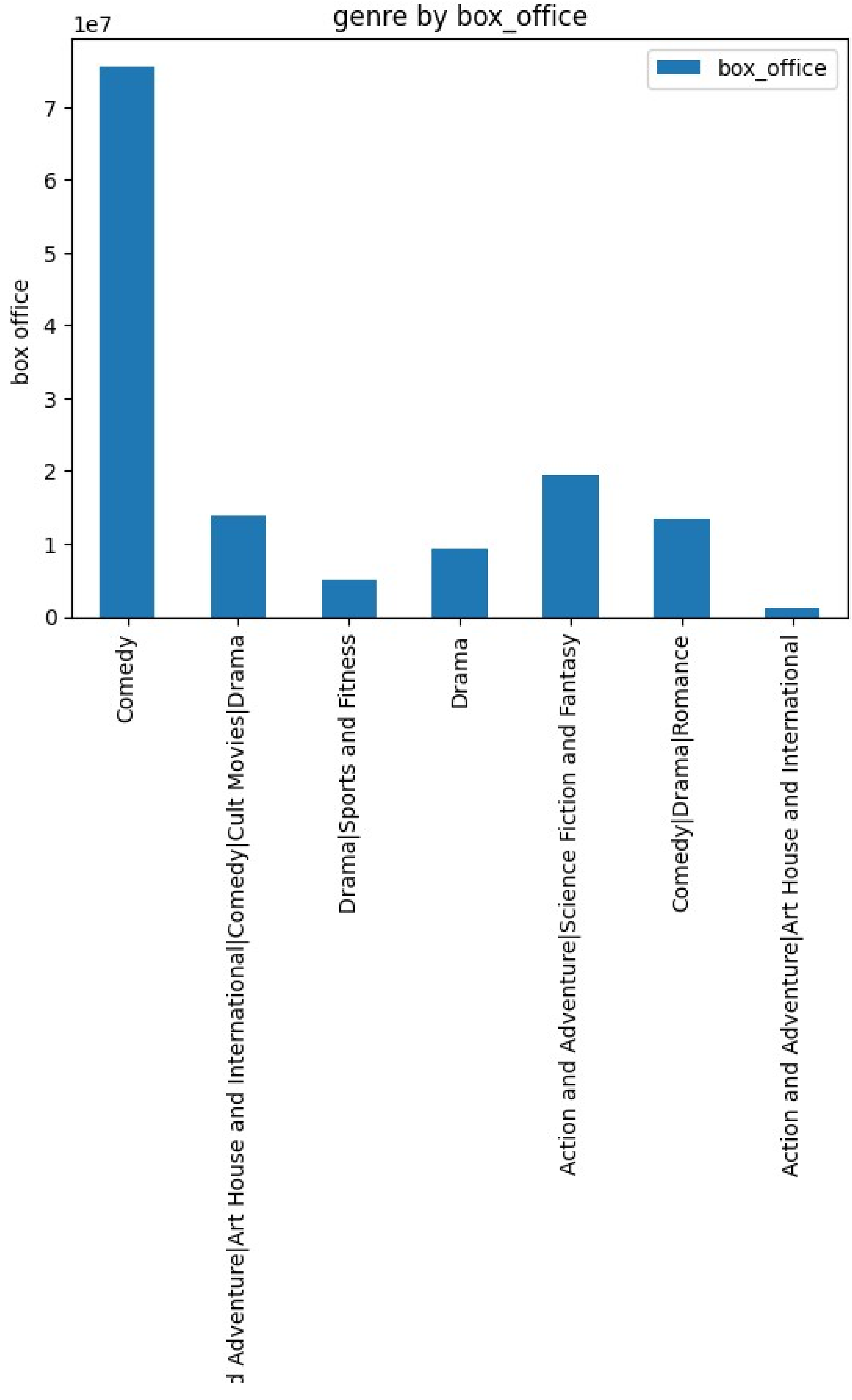
```
0     28.005
1     22.698
2     15.799
6     30.579
12    16.064
Name: popularity, dtype: float64
```

```
# Visualizing the most popular genre
ax=movie_df.plot.bar(
    x= "genre",
    y= "popularity",
    figsize= (10,8)
)
ax.set_title("Popularity by genre")
ax.set_xlabel("genre")
ax.set_ylabel("popularity")
plt.show()
```


Popularity by genre
popularity
popularity
0
5
10
15
20
25
30
Comedy
and Adventure|Art House and International|Comedy|Cult Movies|Drama
Drama|Sports and Fitness
Drama
Action and Adventure|Science Fiction and Fantasy
Comedy|Drama|Romance
Action and Adventure|Art House and International

Drama genre is the most popular genre followed by Comedy genre.

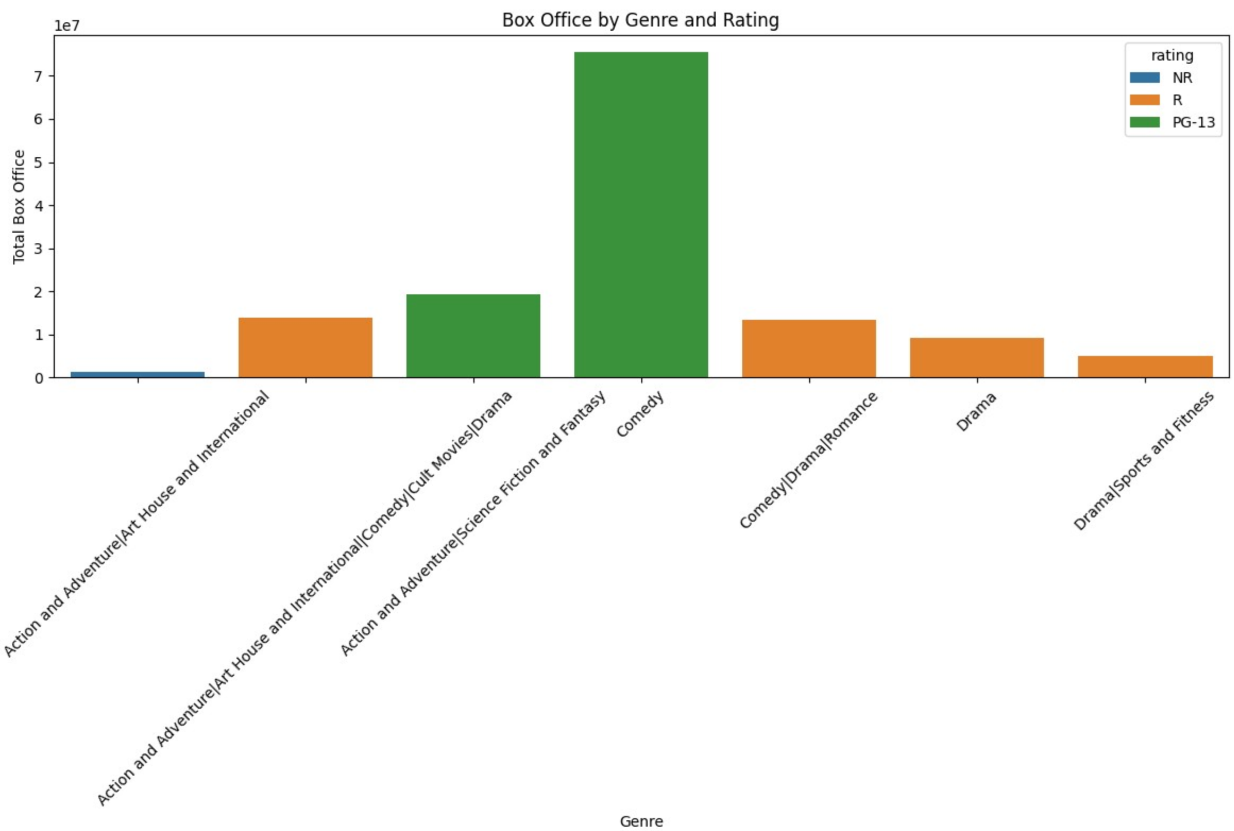
```
# plotting
ax=movie_df.plot.bar(
             x="genre",
             y="box_office",
             rot=90
           )
ax.set_title("genre by box_office")
ax.set_xlabel("genre")
ax.set_ylabel("box office ")
plt.show()
```


genre by box_office
1e7
box_office
box office
0
1
2
3
4
5
6
7
Comedy
d Adventure|Art House and International|Comedy|Cult Movies|Drama
Drama|Sports and Fitness
Drama
Action and Adventure|Science Fiction and Fantasy
Comedy|Drama|Romance
Action and Adventure|Art House and International

Comedy genre generates the highest income

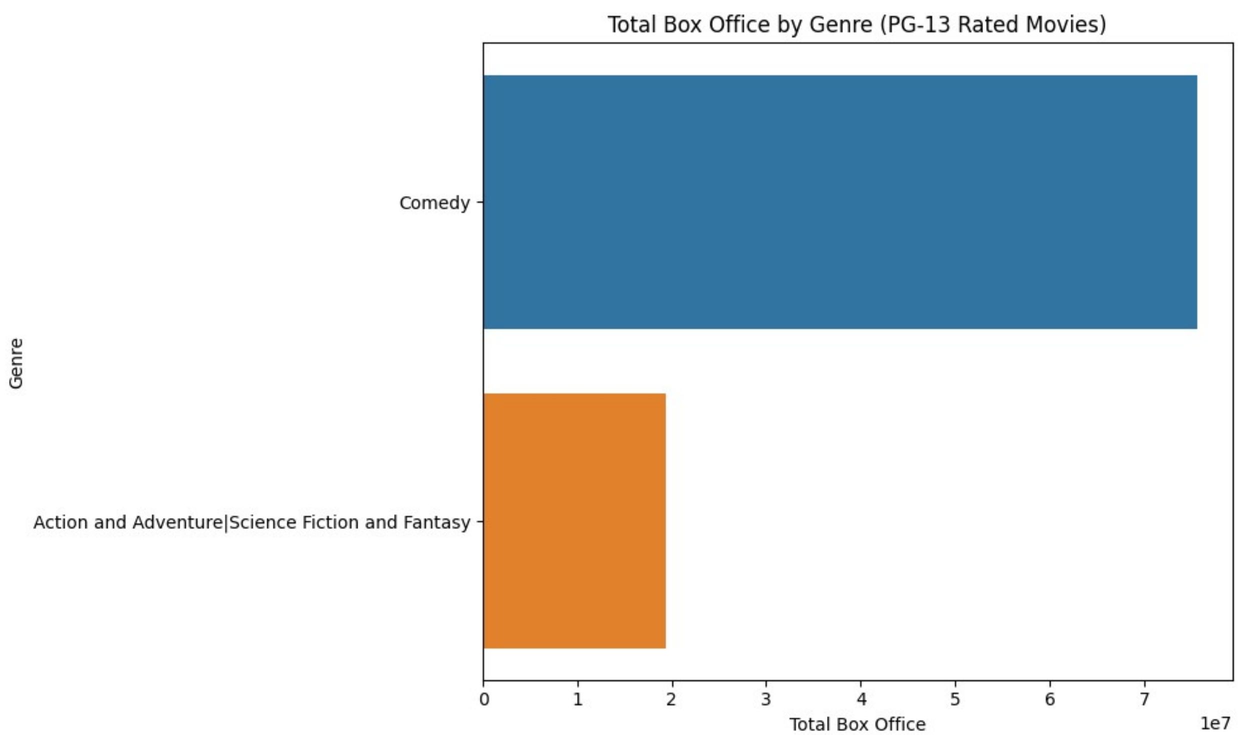
Multivariate Analysis

```
# Groupping by genre and rating, box office
genre_rating_box = (
    movie_df.groupby(["genre", "rating"])["box_office"]
    .sum()
    .reset_index()
)

# Plotting
plt.figure(figsize=(12, 8))
sns.barplot(data=genre_rating_box, x="genre", y="box_office",
hue="rating")

plt.title("Box Office by Genre and Rating")
plt.xlabel("Genre")
plt.ylabel("Total Box Office")
plt.xticks(rotation=45)
plt.tight_layout()
plt.show()
```



Films rated PG-13 tickets sale is high generating more revenue to a movie studio.

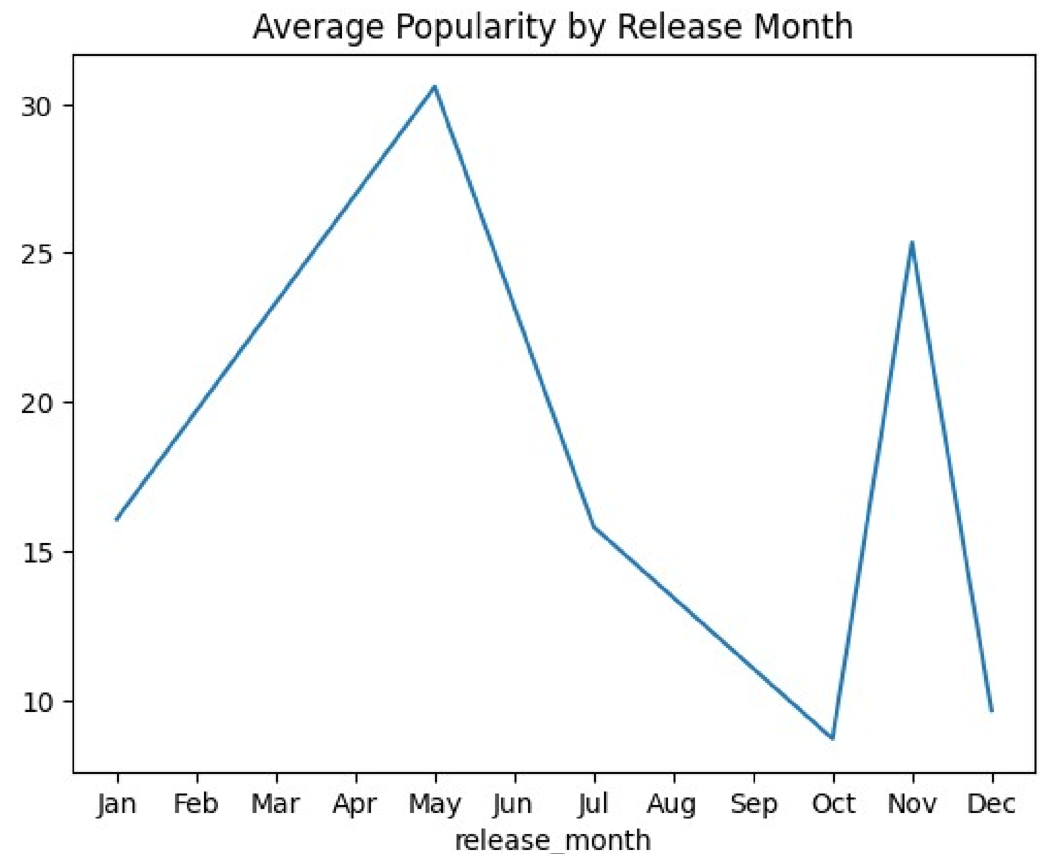
```python
# Filtering PG-13 movies
pg13_df = movie_df[movie_df["rating"] == "PG-13"]

# Group by genre and sum box office
pg13_genre_box = pg13_df.groupby("genre")
["box_office"].sum().reset_index().sort_values(by="box_office",
ascending=False)

# visualizing
plt.figure(figsize=(10, 6))
sns.barplot(data=pg13_genre_box,
            x="box_office",
            y="genre",
            hue="genre",

)
plt.title("Total Box Office by Genre (PG-13 Rated Movies)")
plt.xlabel("Total Box Office")
plt.ylabel("Genre")
plt.tight_layout()
plt.show()
```



Comedy has PG_13 rating and generates more revenue.

```
# Identifying the best time of the month to release a film
movie_df['release_month'] =
pd.to_datetime(movie_df['release_date']).dt.month
monthly_popularity = movie_df.groupby('release_month')
['popularity'].mean()
monthly_popularity.plot()
plt.xticks(range(1,13),
['Jan','Feb','Mar','Apr','May','Jun','Jul','Aug','Sep','Oct','Nov','De
c'])
plt.title('Average Popularity by Release Month')
```

```
Text(0.5, 1.0, 'Average Popularity by Release Month')
```



Best times of the month to release a movie would be from march to may and october to november.

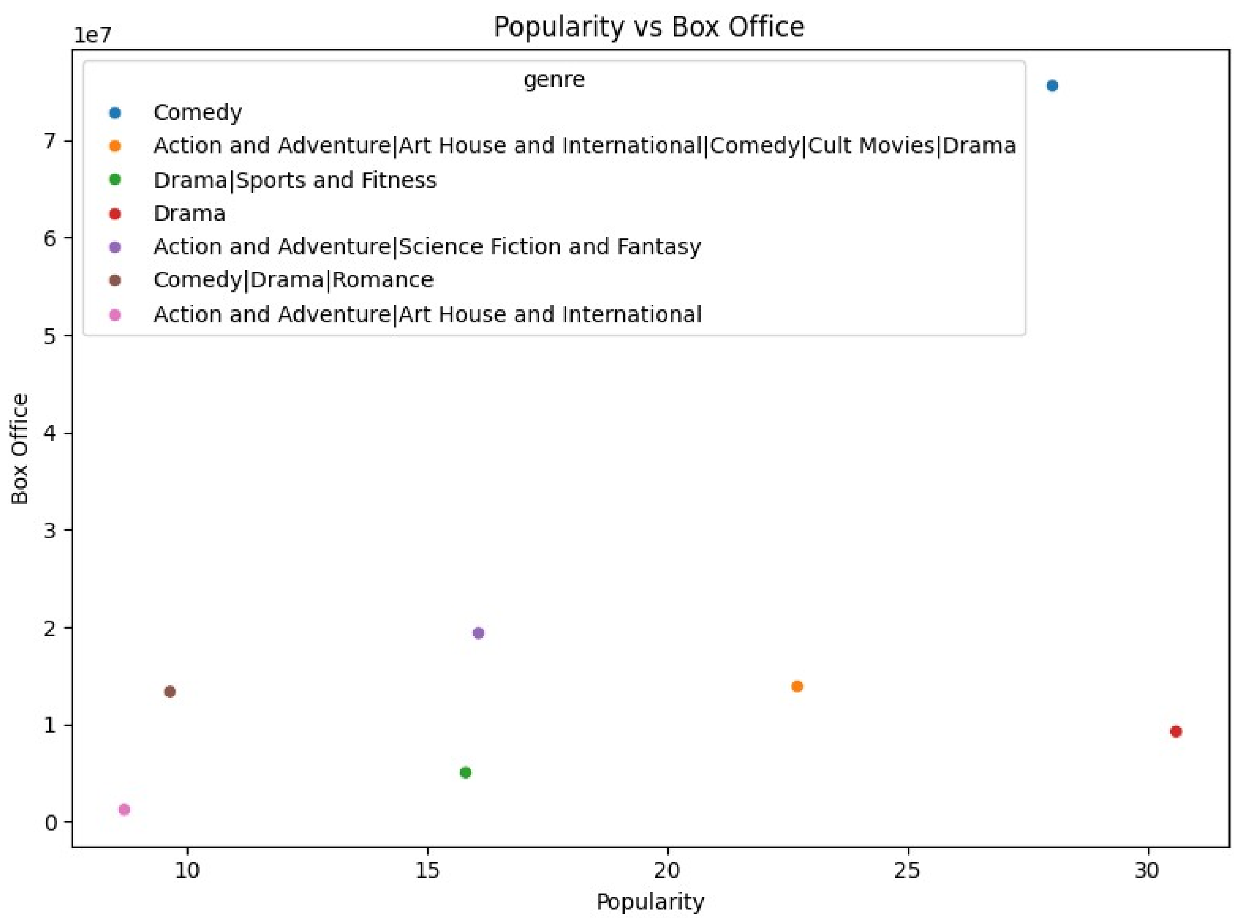
```
movie_df["original_language"]
```

```
0      en
1      en
2      en
6      en
12     en
```

```
20    en
31    sv
Name: original_language, dtype: object
```

English is the most popular language and it generates higher revenue.

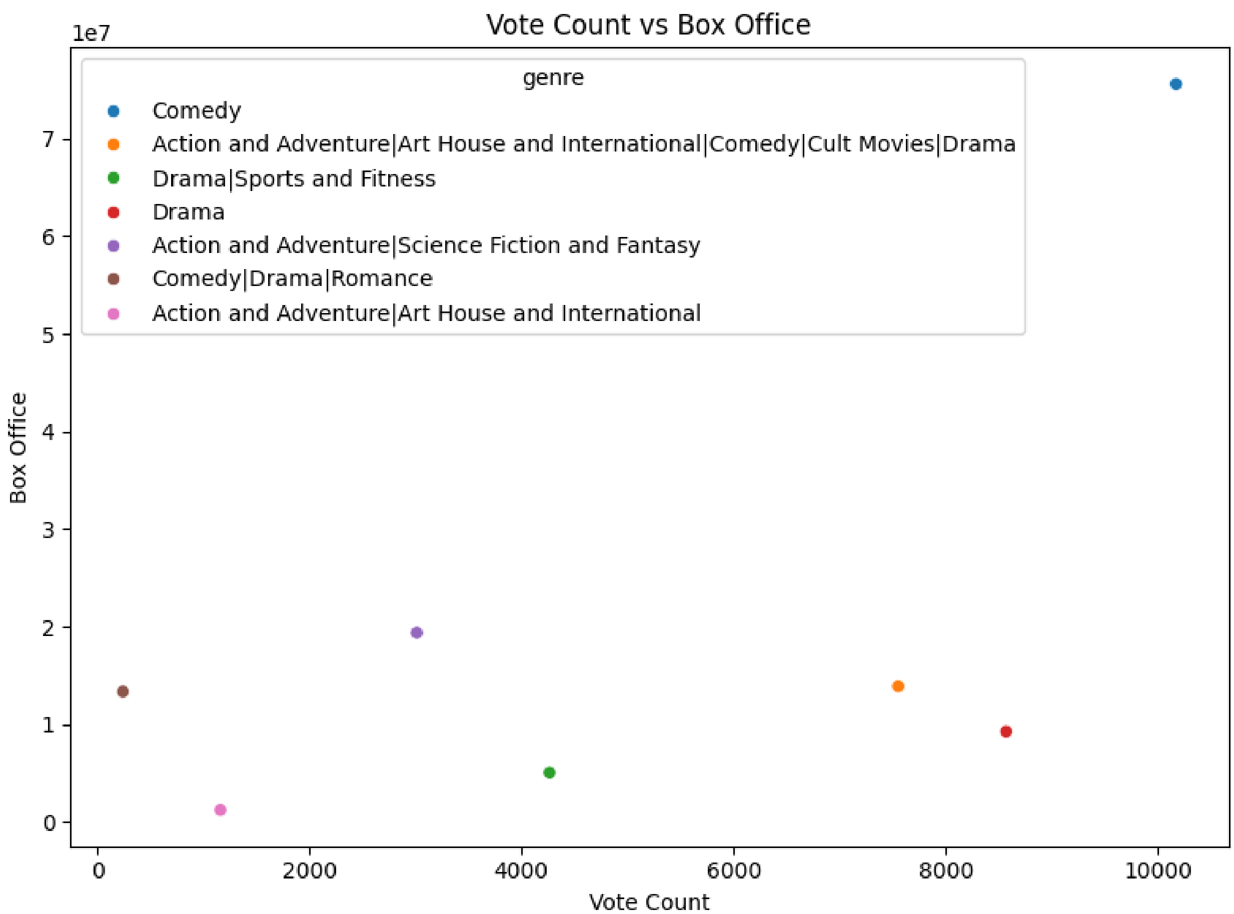
```
# Visualizing the popular genre that generates the highest revenue
plt.figure(figsize=(8, 6))
sns.scatterplot(data=movie_df, x="popularity", y="box_office",
hue="genre")
plt.title("Popularity vs Box Office")
plt.xlabel("Popularity")
plt.ylabel("Box Office")
plt.tight_layout()
plt.show()
```



As much as Drama is the most popular genre, comedy generates the highest revenue

```
# Visualizing the genre with the highest votes that generates the
highest income
plt.figure(figsize=(8, 6))
```

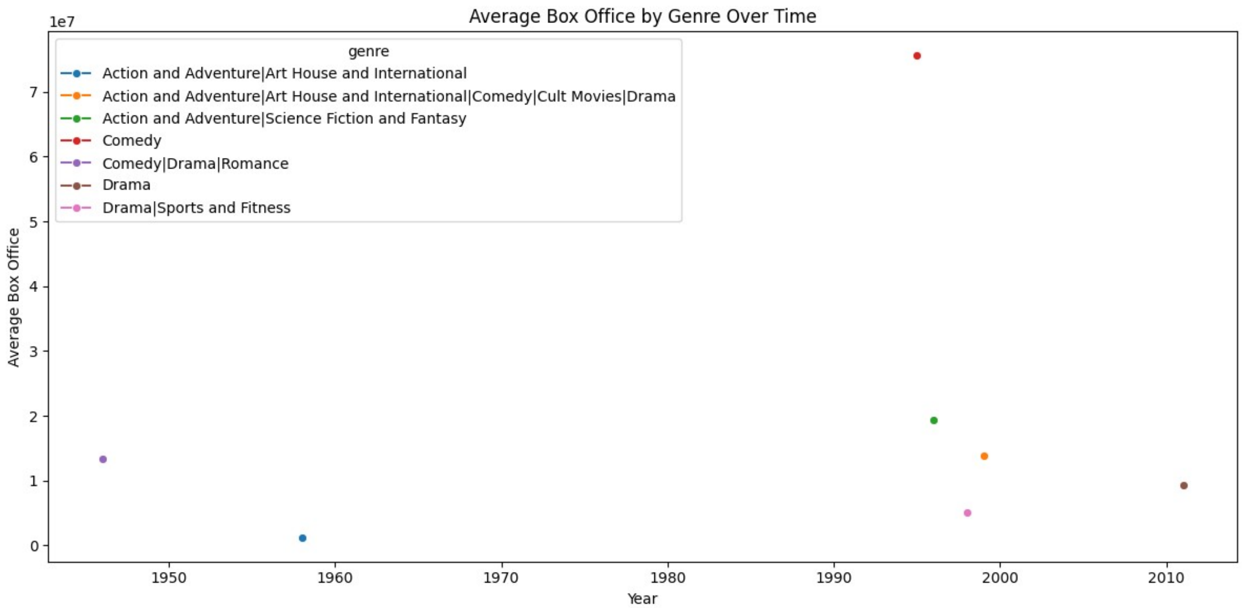
```
sns.scatterplot(data=movie_df, x="vote_count", y="box_office",
hue="genre")
plt.title("Vote Count vs Box Office")
plt.xlabel("Vote Count")
plt.ylabel("Box Office")
plt.tight_layout()
plt.show()
```



Comedy genre has the highest engagement and it generates the highest revenue

```
# Grouping by genre and year, and box office
genre_year_box = (
    movie_df.groupby(["genre", "year"])["box_office"]
    .mean()
    .reset_index()
)

# Plotting
plt.figure(figsize=(12, 6))
sns.lineplot(data=genre_year_box, x="year", y="box_office",
hue="genre", marker="o")
```

```
plt.title("Average Box Office by Genre Over Time")
plt.xlabel("Year")
plt.ylabel("Average Box Office")

plt.tight_layout()
plt.show()
```



Comedy genre has generated the highest revenue over the years.

Conclusion

1. All Genres are equally distributed.
2. Average box office trend over the years has been a little bit sloppy.
3. Comedy genre has the highest number of votes and has generated the highest revenue overtime.
4. Drama is the most popular genre.
5. Comedy generates the highest return on investment.
6. Films rated PG-13 ticket sale is high generating move revenue to a movie studio.
7. Best times of the month to release a movie would be from march to may and october to november.

Recommendations

1. Prioritize Drama and Comedy genres for film production as they are the most popular and generate more revenue than other genres.

2. Maybe you should launch the movie studio with Drama or comedy Film rated PG-13 as they have proven to be popular and comedy has high engagement and a long-term revenue potential.
3. Release the film around may to march months of the year.