```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
```

`Columns Descriptions`

1. Release_Date: Date when the movie was released.
2. Title: Name of the movie.
3. Overview: Brief summary of the movie.
4. Popularity: It is a very important metric computed by TMDB developers based on the number of views per day, votes per day, number of users marked it as "favorite" and "watchlist" for the data, release date and more other metrics.
5. Vote_Count: Total votes received from the viewers.
6. Vote_Average: Average rating based on vote count and the number of viewers out of 10.
7. Original_Language: Original language of the movies. Dubbed version is not considered to be original language.
8. Genre: Categories the movie it can be classified as.
9. Poster_Url: Url of the movie poster.

- `EDA Questions`

1. Q1: What is the most frequent genre in the dataset?
2. Q2: What genres has highest votes?
3. Q3: What movie got the highest popularity? what's its genre?
4. Q4: Which year has the most filmmed movies?

```
from google.colab import files
uploaded = files.upload()
```

Choose Files mymoviedb.csv

- **mymoviedb.csv**(text/csv) - 4208091 bytes, last modified: 9/11/2024 - 100% done

```
Saving mymoviedb.csv to mymoviedb.csv
```

```
data = pd.read_csv('mymoviedb.csv',lineterminator='\n')
```

```
data.head(2)
```

	Release_Date	Title	Overview	Popularity	Vote_Count	Vote_Average	Original_La
0	2021-12-15	Spider-Man: No Way	Peter Parker is unmasked and no longer	5083.954	8940	8.3	

Next steps: Generate code with data | View recommended plots | New interactive sheet

```
data.shape
```

```
(9827, 9)
```

```
data.size
```

```
88443
```

```
data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 9827 entries, 0 to 9826
Data columns (total 9 columns):
 #   Column             Non-Null Count  Dtype  
---  ------             --------------  -----  
 0   Release_Date       9827 non-null   object 
 1   Title              9827 non-null   object 
 2   Overview           9827 non-null   object 
 3   Popularity         9827 non-null   float64
 4   Vote_Count         9827 non-null   int64  
 5   Vote_Average       9827 non-null   float64
 6   Original_Language  9827 non-null   object 
 7   Genre              9827 non-null   object 
 8   Poster_Url         9827 non-null   object 
dtypes: float64(2), int64(1), object(6)
memory usage: 691.1+ KB
```

```
data.describe()
```

	Popularity	Vote_Count	Vote_Average
count	9827.000000	9827.000000	9827.000000
mean	40.326088	1392.805536	6.439534
std	108.873998	2611.206907	1.129759
min	13.354000	0.000000	0.000000
25%	16.128500	146.000000	5.900000
50%	21.199000	444.000000	6.500000
75%	35.191500	1376.000000	7.100000
max	5083.954000	31077.000000	10.000000

```
data.isnull().sum()
```

	0
Release_Date	0
Title	0
Overview	0
Popularity	0
Vote_Count	0
Vote_Average	0
Original_Language	0
Genre	0
Poster_Url	0

dtype: int64

All of them have o null value so no need of null value handling

```
num_col = data.select_dtypes(include=np.number).columns
num_col
```

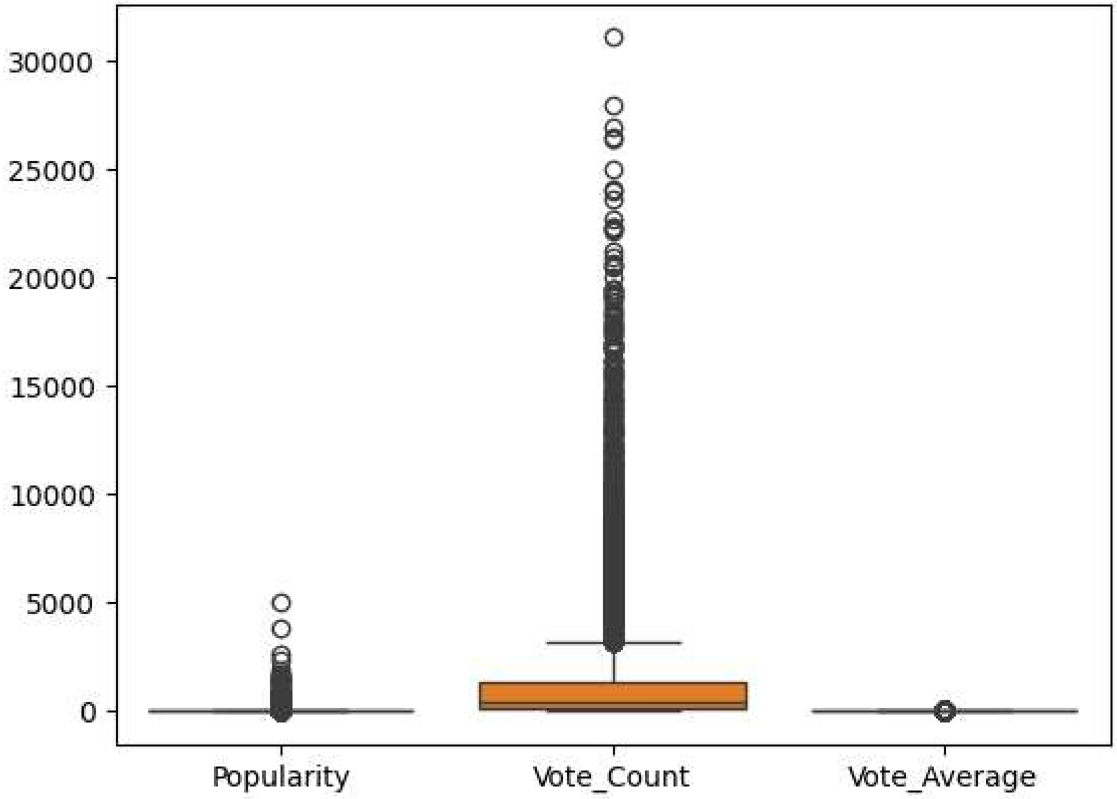
```
Index(['Popularity', 'Vote_Count', 'Vote_Average'], dtype='object')
```

```
cat_col = data.select_dtypes(include='object').columns
cat_col
```

```
Index(['Release_Date', 'Title', 'Overview', 'Original_Language', 'Genre',
       'Poster_Url'],
      dtype='object')
```

```
sns.boxplot(data[num_col])
```

```
<Axes: >
```



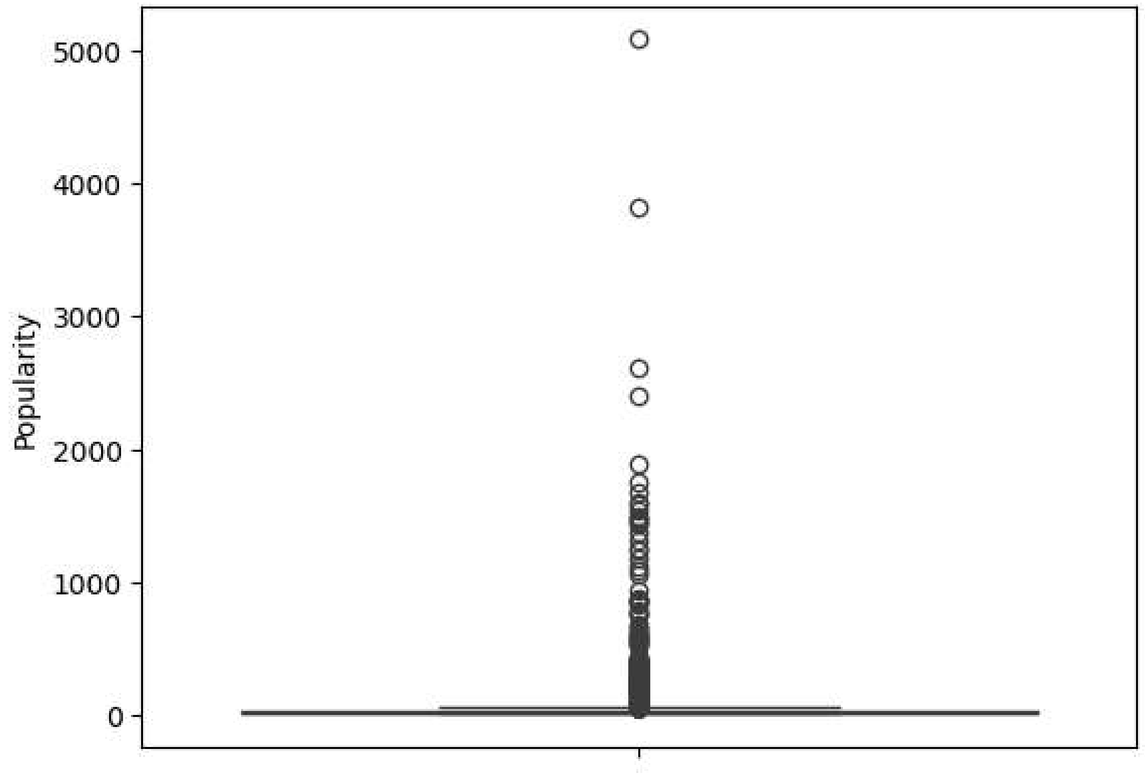
```
data.head()
```

	Release_Date	Title	Overview	Popularity	Vote_Count	Vote_Average	Original_
0	2021-12-15	Spider-Man: No Way Home	Peter Parker is unmasked and no longer able to...	5083.954	8940	8.3	
1	2022-03-01	The Batman	In his second year of fighting crime, Batman u...	3827.658	1151	8.1	
			Stranded at				

Next steps: Generate code with data | View recommended plots | New interactive sheet

```
sns.boxplot(data['Popularity'])
```

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



most of the data came under outlier so no need of this. othervise we doen't have data

lets drop duplicate rows

```
data.head(2)
```

	Release_Date	Title	Overview	Popularity	Vote_Count	Vote_Average	Original_La
0	2021-12-15	Spider-Man: No Way	Peter Parker is unmasked and no longer	5083.954	8940	8.3	

Next steps: Generate code with data | View recommended plots | New interactive sheet

```
data.duplicated().sum()
```

```
0
```

lets drop unwanted columns

i felt Overview and poster url unwanted so i dropped those

```
data.drop(['Overview','Poster_Url'],axis=1,inplace=True)
```

```
data.head()
```

	Release_Date	Title	Popularity	Vote_Count	Vote_Average	Original_Language	
0	2021-12-15	Spider-Man: No Way Home	5083.954	8940	8.3	en	Ad
1	2022-03-01	The Batman	3827.658	1151	8.1	en	I

Next steps: Generate code with data View recommended plots New interactive sheet

```
data['Genre'] = data['Genre'].str.split(', ')
```

```
data.head()
```

	Release_Date	Title	Popularity	Vote_Count	Vote_Average	Original_Language	
0	2021-12-15	Spider-Man: No Way Home	5083.954	8940	8.3	en	Ad
1	2022-03-01	The Batman	3827.658	1151	8.1	en	

Next steps: Generate code with data View recommended plots New interactive sheet

```
data['Genre'] = data['Genre'].explode().reset_index(drop=True)
```

```
data['Rated'] = np.where(data['Vote_Average'] > 7, 'hit', np.where(data['Vote_Average'] >
```

```
data.head()
```

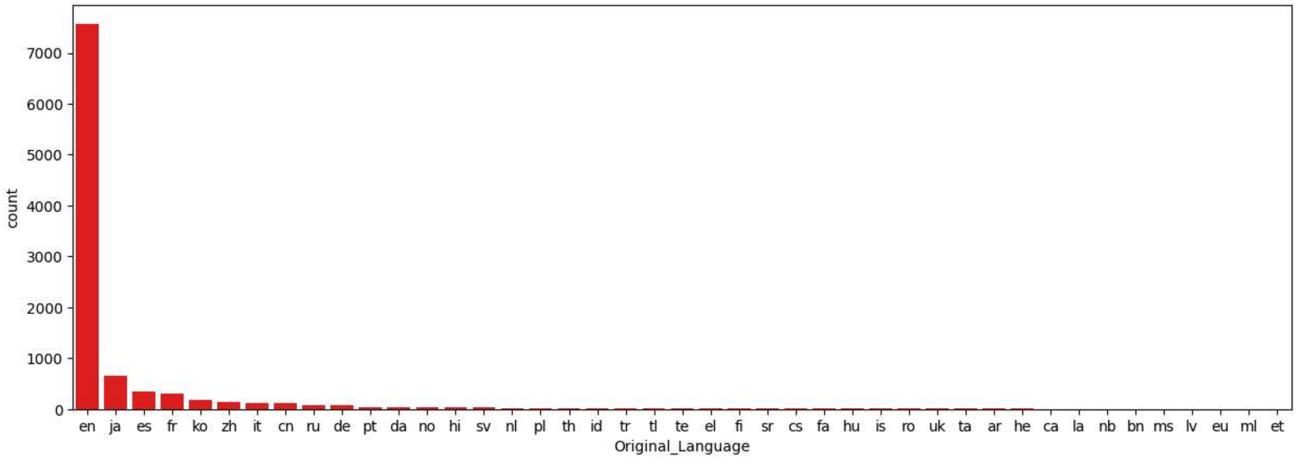
	Release_Date	Title	Popularity	Vote_Count	Vote_Average	Original_Language	
0	2021-12-15	Spider-Man: No Way Home	5083.954	8940	8.3	en	
1	2022-03-01	The Batman	3827.658	1151	8.1	en	Ad

Next steps: Generate code with data | View recommended plots | New interactive sheet

Data Visulization

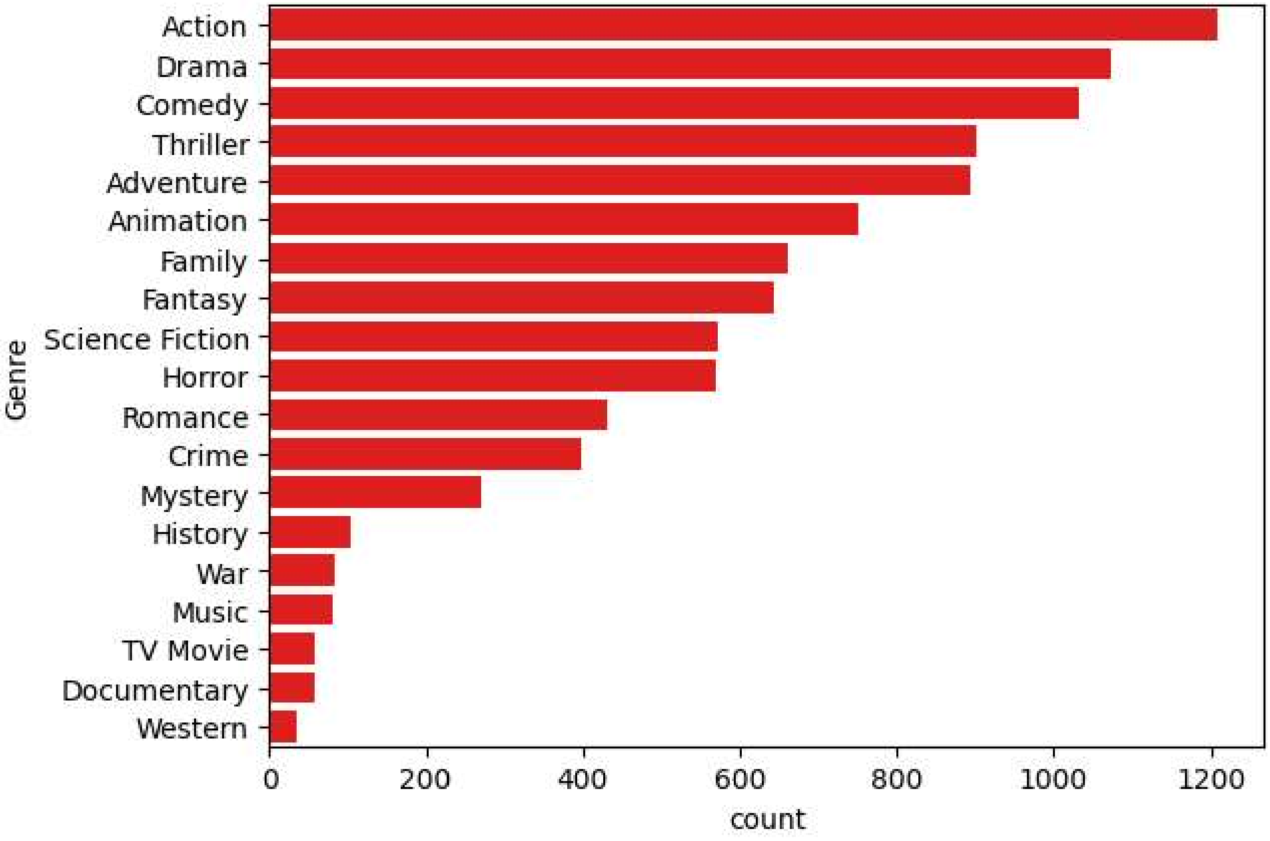
```
plt.figure(figsize=(15,5))
sns.barplot(x=data['Original_Language'].value_counts().index,y=data['Original_Language'].
```

```
<Axes: xlabel='Original_Language', ylabel='count'>
```



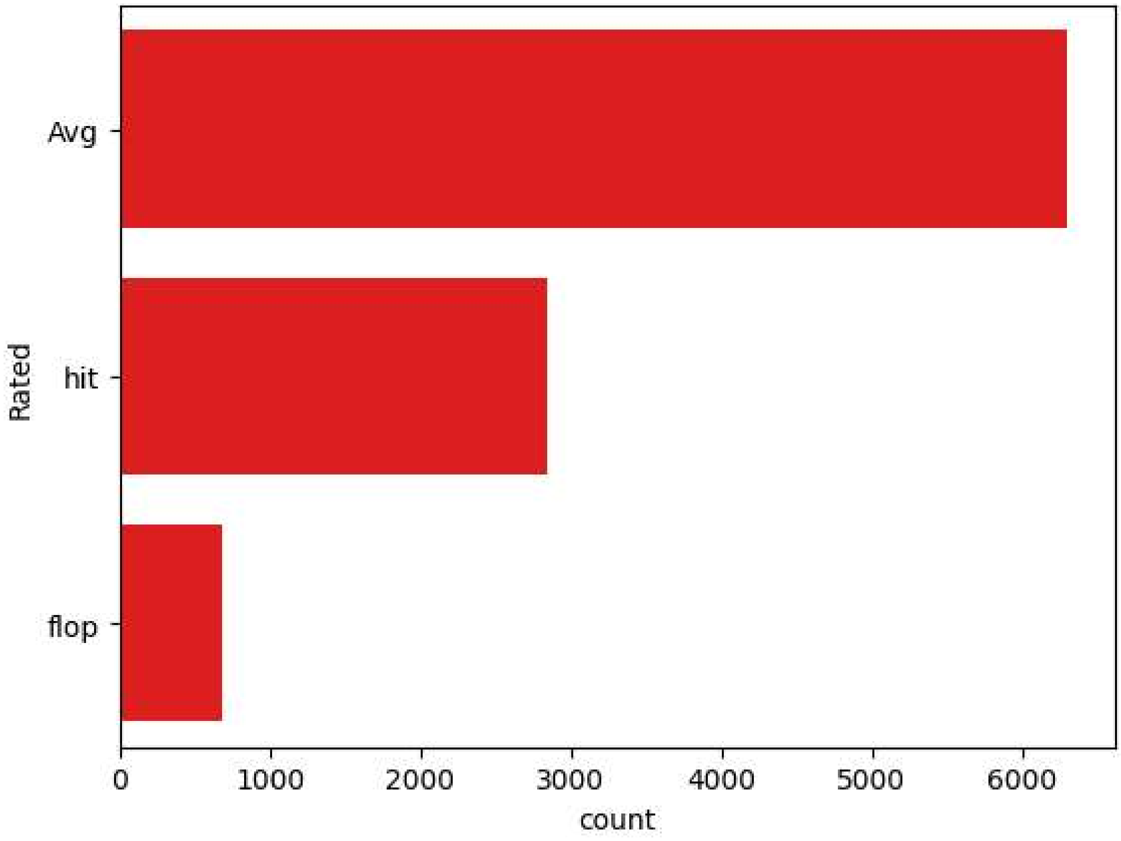
Q1: What is the most frequent genre in the dataset?

```
sns.barplot(y=data['Genre'].value_counts().index,x=data['Genre'].value_counts(),color='re
<Axes: xlabel='count', ylabel='Genre'>
```



```
sns.barplot(y=data['Rated'].value_counts().index,x=data['Rated'].value_counts(),color='re
```

```
<Axes: xlabel='count', ylabel='Rated'>
```



Q2: What genres has highest votes?

```
data.head(2)
```

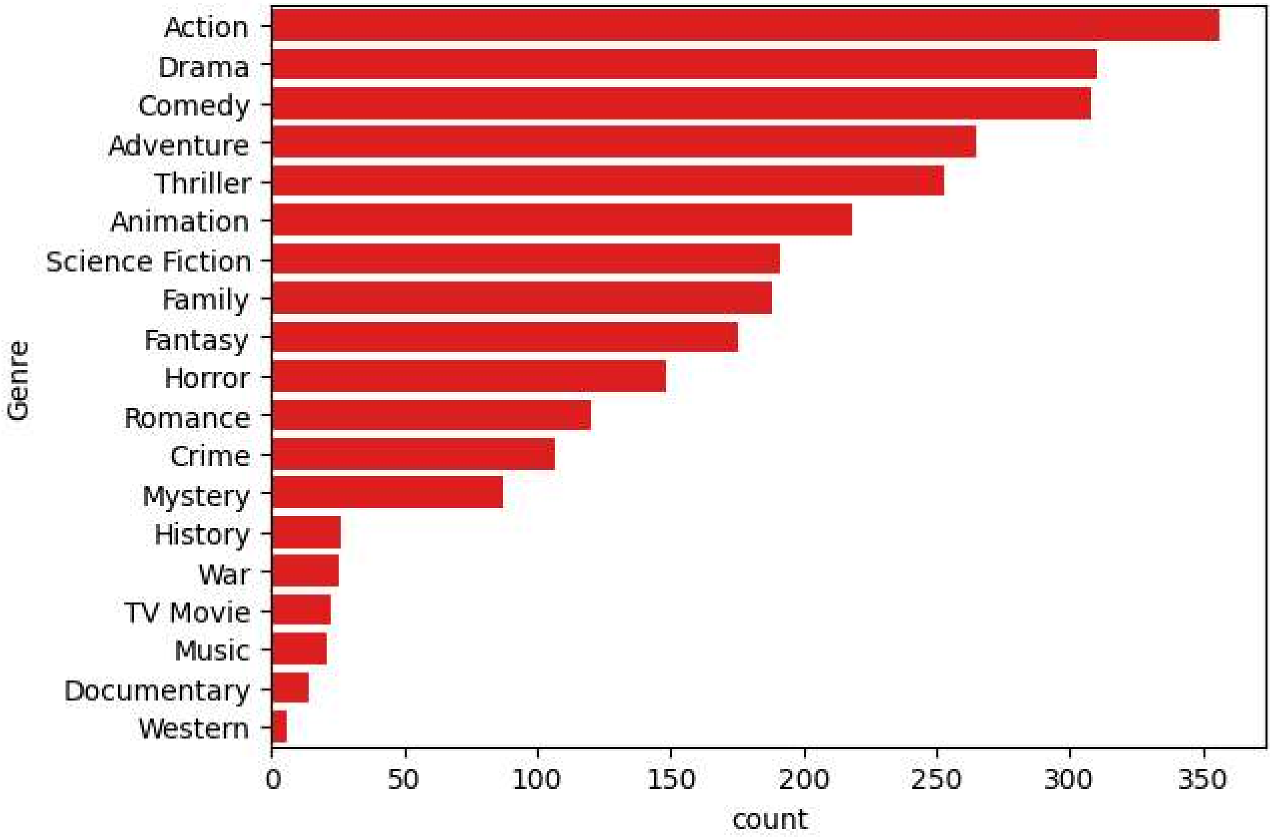
	Release_Date	Title	Popularity	Vote_Count	Vote_Average	Original_Language
0	2021-12-15	Spider-Man: No	5083.954	8940	8.3	en

Next steps: Generate code with data | View recommended plots | New interactive sheet

```
popular_movies = data[data['Rated'] == 'hit']
```

```
sns.barplot(y=popular_movies['Genre'].value_counts().index,x=popular_movies['Genre'].valu
```

```
<Axes: xlabel='count', ylabel='Genre'>
```



Q3: What movie got the highest popularity? what's its genre?

```
data[data['Popularity'] == data['Popularity'].max()]
```

	Release_Date	Title	Popularity	Vote_Count	Vote_Average	Original_Language	Gen
		Spider-					

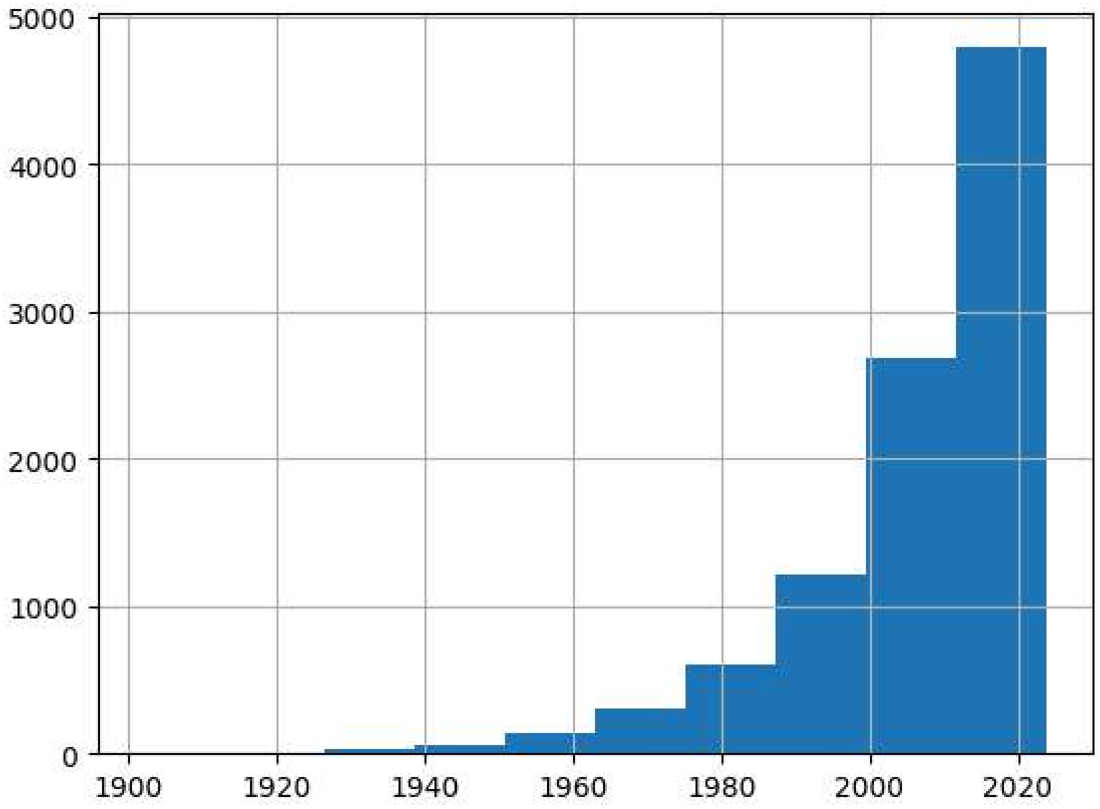
Q4: Which year has the most filmmed movies?

```
data['Release_Date'] = pd.to_datetime(data['Release_Date'])
```

```
data['Release_year'] = data['Release_Date'].dt.year
```

```
data['Release_year'].hist()
```

```
<Axes: >
```



Start coding or generate with AI.

```
                 count
Release_year
    1902             1
    1920             1
    1921             2
    1922             2
    1925             1
```