


Start coding or [generate](#) with AI.

Flix_Focus (NCER) Prajwal, Kalpesh


Start coding or [generate](#) with AI.

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


```
df = pd.read_csv('/content/mymoviedb.csv', lineterminator='\n')
df.head()
```

	Release_Date	Title	Overview	Popularity	Vote_Count	Vote_Average	Original_Language	Genre	Poster_Url
0	2021-12-15	Spider-Man: No Way Home	Peter Parker is unmasked and no longer able to...	5083.954	8940	8.3	en	Action, Adventure, Science Fiction	https://image.tmdb.org/t/p/original/1g0dhYtq4i...
1	2022-03-01	The Batman	In his second year of fighting crime, Batman u...	3827.658	1151	8.1	en	Crime, Mystery, Thriller	https://image.tmdb.org/t/p/original/74xTEgt7R3...
2	2022-02-25	No Exit	Stranded at a rest stop in the mountains durin...	2618.087	122	6.3	en	Thriller	https://image.tmdb.org/t/p/original/vDHsLnOWKl...
3	2021-11-24	Encanto	The tale of an extraordinary family, the Madri...	2402.201	5076	7.7	en	Animation, Comedy, Family, Fantasy	https://image.tmdb.org/t/p/original/4j0PNHkMr5...
4	2021-12-22	The King's Man	As a collection of history's worst tyrants and...	1895.511	1793	7.0	en	Action, Adventure, Thriller, War	https://image.tmdb.org/t/p/original/aq4Pwv5Xeu...

```
df.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
```


```
# exploring genres column
df['Genre'].head()
```



	Genre
0	Action, Adventure, Science Fiction
1	Crime, Mystery, Thriller
2	Thriller
3	Animation, Comedy, Family, Fantasy
4	Action, Adventure, Thriller, War


df.info()

```
# check for duplicated rows
df.duplicated().sum()
```



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

```
# exploring summary statistics
df.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 cleaning

```
df.head()
```

	Release_Date	Title	Overview	Popularity	Vote_Count	Vote_Average	Original_Language	Genre	Poster_Url
0	2021-12-15	Spider-Man: No Way Home	Peter Parker is unmasked and no longer able to...	5083.954	8940	8.3	en	Action, Adventure, Science Fiction	https://image.tmdb.org/t/p/original/1g0dhYtq4i...
1	2022-03-01	The Batman	In his second year of fighting crime, Batman u...	3827.658	1151	8.1	en	Crime, Mystery, Thriller	https://image.tmdb.org/t/p/original/74xTEgt7R3...
2	2022-02-25	No Exit	Stranded at a rest stop in the mountains durin...	2618.087	122	6.3	en	Thriller	https://image.tmdb.org/t/p/original/vDHsLnOWKl...
3	2021-11-24	Encanto	The tale of an extraordinary family, the Madri...	2402.201	5076	7.7	en	Animation, Comedy, Family, Fantasy	https://image.tmdb.org/t/p/original/4j0PNHkMr5...
4	2021-12-22	The King's Man	As a collection of history's	1805.511	1702	7.0	en	Action, Adventure,	https://image.tmdb.org/t/p/original/1g0dhYtq4i...

```
# casting column a
df['Release_Date'] = pd.to_datetime(df['Release_Date'])
# confirming changes
print(df['Release_Date'].dtypes)
```

```
datetime64[ns]
```

```
df['Release_Date'] = df['Release_Date'].dt.year
df['Release_Date'].dtypes
```

```
dtype('int32')
```

```
df.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   int32
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), int32(1), int64(1), object(5)
memory usage: 652.7+ KB
```

```
df.head()
```



	Release_Date	Title	Overview	Popularity	Vote_Count	Vote_Average	Original_Language	Genre	Poster_Url
0	2021	Spider-Man: No Way Home	Peter Parker is unmasked and no longer able to...	5083.954	8940	8.3	en	Action, Adventure, Science Fiction	https://image.tmdb.org/t/p/original/1g0dhYtq4i...
1	2022	The Batman	In his second year of fighting crime, Batman u...	3827.658	1151	8.1	en	Crime, Mystery, Thriller	https://image.tmdb.org/t/p/original/74xTEgt7R3...
2	2022	No Exit	Stranded at a rest stop in the mountains durin...	2618.087	122	6.3	en	Thriller	https://image.tmdb.org/t/p/original/vDHsLnOWKl...
3	2021	Encanto	The tale of an extraordinary family, the Madri...	2402.201	5076	7.7	en	Animation, Comedy, Family, Fantasy	https://image.tmdb.org/t/p/original/4j0PNHkMr5...
4	2021	The King's Man	As a collection of history's	1805.511	1702	7.0	en	Action, Adventure,	https://image.tmdb.org/t/p/original/1g0dhYtq4i...

```
# making list of column to be dropped
cols = ['Overview', 'Original_Language', 'Poster_Url']
# making list of column to be dropped
cols = ['Overview', 'Original_Language', 'Poster_Url']
```

```
df.head()
```



	Release_Date	Title	Overview	Popularity	Vote_Count	Vote_Average	Original_Language	Genre	Poster_Url
0	2021	Spider-Man: No Way Home	Peter Parker is unmasked and no longer able to...	5083.954	8940	8.3	en	Action, Adventure, Science Fiction	https://image.tmdb.org/t/p/original/1g0dhYtq4i...
1	2022	The Batman	In his second year of fighting crime, Batman u...	3827.658	1151	8.1	en	Crime, Mystery, Thriller	https://image.tmdb.org/t/p/original/74xTEgt7R3...
2	2022	No Exit	Stranded at a rest stop in the mountains durin...	2618.087	122	6.3	en	Thriller	https://image.tmdb.org/t/p/original/vDHsLnOWKl...
3	2021	Encanto	The tale of an extraordinary family, the Madri...	2402.201	5076	7.7	en	Animation, Comedy, Family, Fantasy	https://image.tmdb.org/t/p/original/4j0PNHkMr5...
4	2021	The King's Man	As a collection of history's	1805.511	1702	7.0	en	Action, Adventure,	https://image.tmdb.org/t/p/original/1g0dhYtq4i...

```
def catigorize_col (df, col, labels):
    """
    catigorizes a certain column based on its quartiles

    Args:
    (df) df - dataframe we are proccesing
    (col) str - to be catigorized column's name
    (labels) list - list of labels from min to max
```

Returns:

```
(df) df - dataframe with the categorized col
"""
```

```
# setting the edges to cut the column accordingly
edges = [df[col].describe()['min'],
df[col].describe()['25%'],
df[col].describe()['50%'],
df[col].describe()['75%'],
df[col].describe()['max']]
df[col] = pd.cut(df[col], edges, labels = labels, duplicates='drop')
return df
```

```
# define labels for edges
```

```
labels = ['not_popular', 'below_avg', 'average', 'popular']
```

```
# categorize column based on labels and edges
```

```
categorize_col(df, 'Vote_Average', labels)
```

```
# confirming changes
```

```
df['Vote_Average'].unique()
```

```
→ ['popular', 'below_avg', 'average', 'not_popular', NaN]
Categories (4, object): ['not_popular' < 'below_avg' < 'average' < 'popular']
```

```
# exploring column
```

```
df['Vote_Average'].value_counts()
```

```
→
```

	count
Vote_Average	
not_popular	2467
popular	2450
average	2412
below_avg	2398

df.dropna(inplace=True)

```
# dropping NaNs
```

```
df.dropna(inplace = True)
```

```
# confirming
```

```
df.isna().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

df.info()

df.head()



	Release_Date	Title	Overview	Popularity	Vote_Count	Vote_Average	Original_Language	Genre	Poster_Url
0	2021	Spider-Man: No Way Home	Peter Parker is unmasked and no longer able to...	5083.954	8940	popular	en	Action, Adventure, Science Fiction	https://image.tmbd.org/t/p/original/1g0dhYtq4i...
1	2022	The Batman	In his second year of fighting crime, Batman u...	3827.658	1151	popular	en	Crime, Mystery, Thriller	https://image.tmbd.org/t/p/original/74xTEgt7R3...
2	2022	No Exit	Stranded at a rest stop in the mountains durin...	2618.087	122	below_avg	en	Thriller	https://image.tmbd.org/t/p/original/vDHsLnOWKI...
3	2021	Encanto	The tale of an extraordinary family, the Madri...	2402.201	5076	popular	en	Animation, Comedy, Family, Fantasy	https://image.tmbd.org/t/p/original/4j0PNHkMr5...
4	2021	The King's Man	As a collection of history's	1805.511	1702	average	en	Action, Adventure,	https://image.tmbd.org/t/p/original/1g0dhYtq4i...

- we'd split genres into a list and then explode our dataframe to have only one genre per row for each movie

```
# split the strings into lists
df['Genre'] = df['Genre'].str.split(',')
# explode the lists
df = df.explode('Genre').reset_index(drop=True)
df.head()
```



	Release_Date	Title	Overview	Popularity	Vote_Count	Vote_Average	Original_Language	Genre	Poster_Url
0	2021	Spider-Man: No Way Home	Peter Parker is unmasked and no longer able to...	5083.954	8940	popular	en	Action	https://image.tmdb.org/t/p/original/1g0dhYtq4i...
1	2021	Spider-Man: No Way Home	Peter Parker is unmasked and no longer able to...	5083.954	8940	popular	en	Adventure	https://image.tmdb.org/t/p/original/1g0dhYtq4i...
2	2021	Spider-Man: No Way Home	Peter Parker is unmasked and no longer able to...	5083.954	8940	popular	en	Science Fiction	https://image.tmdb.org/t/p/original/1g0dhYtq4i...
3	2022	The Batman	In his second year of fighting crime, Batman u...	3827.658	1151	popular	en	Crime	https://image.tmdb.org/t/p/original/74xTEgt7R3...
4	2022	The Batman	In his second year of fighting crime, Batman u...	3827.658	1151	popular	en	Mystery	https://image.tmdb.org/t/p/original/74xTEgt7R3...



```
# casting column into category
df['Genre'] = df['Genre'].astype('category')
# confirming changes
df['Genre'].dtypes
```



```
CategoricalDtype(categories=['Action', 'Adventure', 'Animation', 'Comedy', 'Crime',
                             'Documentary', 'Drama', 'Family', 'Fantasy', 'History',
                             'Horror', 'Music', 'Mystery', 'Romance', 'Science Fiction',
                             'TV Movie', 'Thriller', 'War', 'Western'],
                  ordered=False, categories_dtype=object)
```

```
df.info()
```



```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 25552 entries, 0 to 25551
Data columns (total 9 columns):
#   Column          Non-Null Count  Dtype
---  -
0   Release_Date     25552 non-null  int32
1   Title            25552 non-null  object
2   Overview         25552 non-null  object
3   Popularity       25552 non-null  float64
4   Vote_Count       25552 non-null  int64
5   Vote_Average     25552 non-null  category
6   Original_Language 25552 non-null  object
7   Genre            25552 non-null  category
8   Poster_Url       25552 non-null  object
dtypes: category(2), float64(1), int32(1), int64(1), object(4)
memory usage: 1.3+ MB
```

```
df.nunique()
```



	0
Release_Date	100
Title	9415
Overview	9722
Popularity	8088
Vote_Count	3265
Vote_Average	4
Original_Language	42
Genre	19
Poster_Url	9727

dtype: int64

✓ Data Visualization

here, we'd use Matplotlib and seaborn for making some informative visuals to gain insights about our data.

```
# setting up seaborn configurations
sns.set_style('whitegrid')
```

```
# showing stats. on genre column
df['Genre'].describe()
```



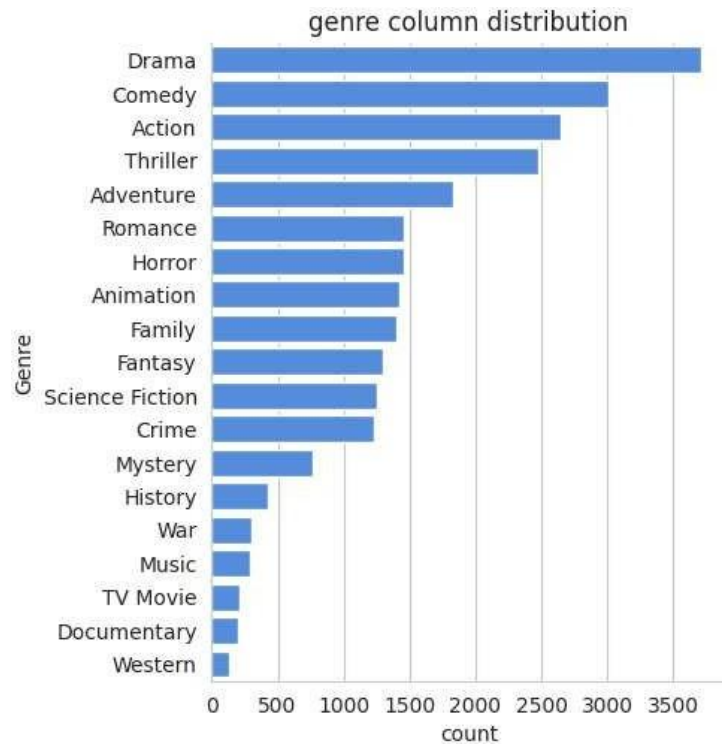
	Genre
count	25552
unique	19
top	Drama
freq	3715

dtype: object

```
# visualizing genre column
sns.catplot(y = 'Genre', data = df, kind = 'count',
            order = df['Genre'].value_counts().index,
            color = '#4287f5')
```

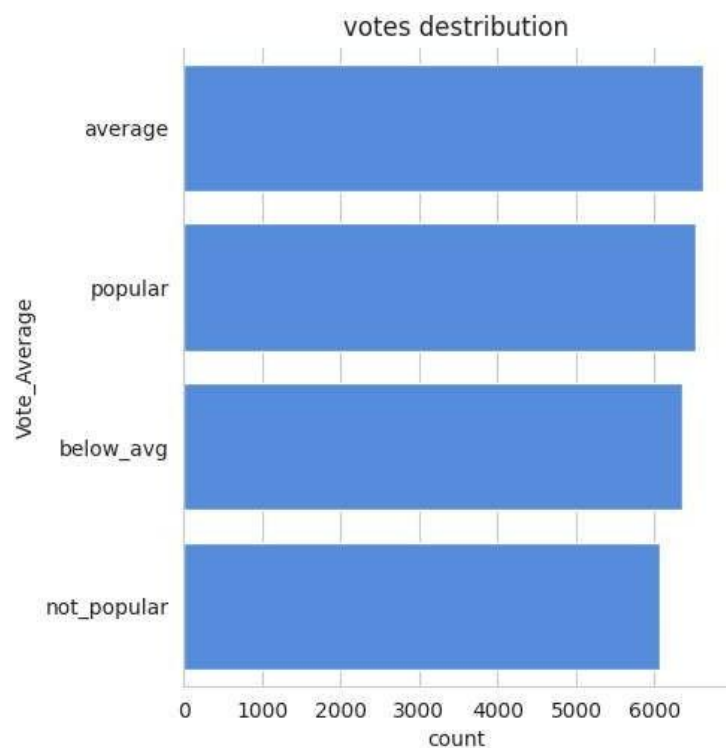


```
plt.title('genre column distribution')
plt.show()
```



we can notice from the above visual that Drama genre is the most frequent genre in our dataset and has appeared more than 14% of the times among 19 other genres.

```
# visualizing vote_average column
sns.catplot(y = 'Vote_Average', data = df, kind = 'count',
            order = df['Vote_Average'].value_counts().index,
            color = '#4287f5')
plt.title('votes distribution')
plt.show()
```



```
# checking max popularity in dataset  
df[df['Popularity'] == df['Popularity'].max()]
```



	Release_Date	Title	Overview	Popularity	Vote_Count	Vote_Average	Original_Language	Genre	Poster_Ur1
0	2021	Spider-Man: No Way Home	Peter Parker is unmasked and no longer able to...	5083.954	8940	popular	en	Action	https://image.tmdb.org/t/p/original/1g0dhYtq4i...
1	2021	Spider-Man: No Way Home	Peter Parker is unmasked and no longer able to...	5083.954	8940	popular	en	Adventure	https://image.tmdb.org/t/p/original/1g0dhYtq4i...
2	2021	Spider-Man: No Way Home	Peter Parker is unmasked and no longer able to...	5083.954	8940	popular	en	Science Fiction	https://image.tmdb.org/t/p/original/1g0dhYtq4i...

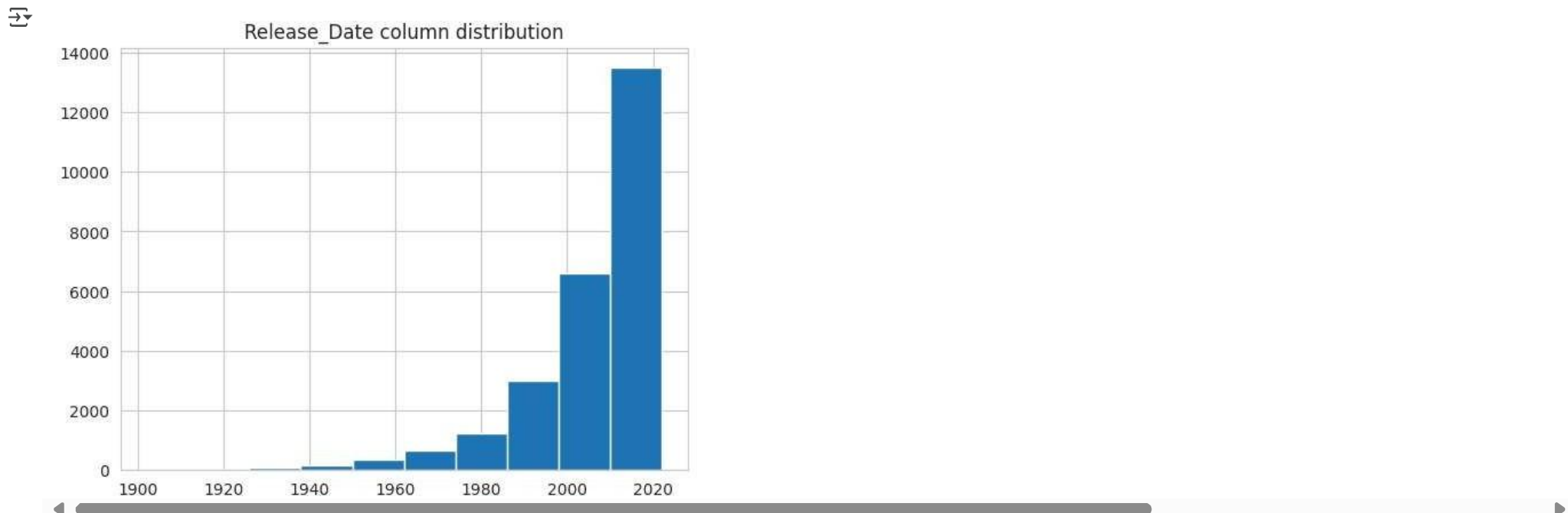
✓ Lowest popularity

```
# checking max popularity in dataset  
df[df['Popularity'] == df['Popularity'].min()]
```

	Release_Date	Title	Overview	Popularity	Vote_Count	Vote_Average	Original_Language	Genre	Poster_Ur1
25546	2021	The United States vs. Billie Holiday	Billie Holiday spent much of her career being ...	13.354	152	average	en	Music	https://image.tmdb.org/t/p/original/vEzkxuE2sJ...
25547	2021	The United States vs. Billie Holiday	Billie Holiday spent much of her career being ...	13.354	152	average	en	Drama	https://image.tmdb.org/t/p/original/vEzkxuE2sJ...
25548	2021	The United States vs. Billie Holiday	Billie Holiday spent much of her career being ...	13.354	152	average	en	History	https://image.tmdb.org/t/p/original/vEzkxuE2sJ...
25549	1984	Threads	Documentary style account of a nuclear holocau...	13.354	186	popular	en	War	https://image.tmdb.org/t/p/original/IBhU4U9Eeh...
25550	1984	Threads	Documentary style account of a nuclear holocau...	13.354	186	popular	en	Drama	https://image.tmdb.org/t/p/original/IBhU4U9Eeh...
25551	1984	Threads	Documentary style account of a nuclear holocau...	13.354	186	popular	en	Science Fiction	https://image.tmdb.org/t/p/original/IBhU4U9Eeh...

✓ Which year has the most filmed movies?

```
df['Release_Date'].hist()
plt.title('Release_Date column distribution')
plt.show()
```



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

Ans :- Drama genre is the most frequent genre in our dataset and has appeared more than 14% of the times among 19 other genres.

Q2: What genres has highest votes ?

Ans :- we have 25.5% of our dataset with popular vote (6520 rows). Drama again gets the highest popularity among fans by being having more than 18.5% of movies popularities.

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

Ans :- Spider-Man: No Way Home has the highest popularity rate in our dataset and it has genres of Action , Adventure and Sience Fiction .

Q4: What movie got the lowest popularity ? what's its genre ?

Ans :- The united states, thread' has the highest lowest rate in our dataset and it has genres of music , drama , 'war', 'sci-fi' and history.

Q4: Which year has the most filmed movies?