In [57]:
```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
```

In [59]:
```
df = pd.read_csv('mymoviedb.csv', lineterminator='\n')
df.head()
```

Out[59]:

	Release_Date	Title	Overview	Popularity	Vote_Count	Vote_Average	Original_Lan
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	
2	2022-02-25	No Exit	Stranded at a rest stop in the mountains durin...	2618.087	122	6.3	
3	2021-11-24	Encanto	The tale of an extraordinary family, the Madri...	2402.201	5076	7.7	
4	2021-12-22	The King's Man	As a collection of history's worst tyrants and...	1895.511	1793	7.0	

In [61]:
```
# viewing dataset info
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
```

- looks like our dataset has **no NaNs!**
- **Overview**, **Original_Language** and **Poster-Url** wouldn't be so useful during analysis
- **Release_Date** column needs to be casted intodate time and to extract only the year value

In [64]:
```
df['Genre'].head()
```

Out[64]:
```
0    Action, Adventure, Science Fiction
1               Crime, Mystery, Thriller
2                               Thriller
3    Animation, Comedy, Family, Fantasy
4      Action, Adventure, Thriller, War
Name: Genre, dtype: object
```

- • **genres** are saperated by commas followed by **whitespaces**.

In [67]:
```
# check for duplicated rows
df.duplicated().sum()
```

Out[67]:
```
0
```

• our dataset has **no duplicated** rows either.

In [70]:
```
# exploring summary statistics
df.describe()
```

Out[70]:

	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

• Exploration Summary

- The dataset contains **9,827** rows and **9** columns.
- The data is clean, with no missing values or duplicates.
- The Release_Date column should be converted to a proper date format, and we can extract just the year if needed.
- Columns like Overview, Original_Language, and Poster_Url are not useful for our analysis and can be ignored.
- There are some outliers in the Popularity column that may affect the analysis.
- The Vote_Average column can be categorized to better understand rating patterns.
- The Genre column contains comma-separated values with extra spaces, which should be cleaned and split properly for accurate analysis.

In [ ]:

Casting **Release_Date** column and extracing year values

In [75]:
```
df['Release_Date'] = pd.to_datetime(df['Release_Date'])

print(df['Release_Date'].dtype)
```

```
datetime64[ns]
```

In [77]:
```
df['Release_Date'] = df['Release_Date'].dt.year

print(df['Release_Date'].dtype)
```

```
int32
```

In [79]:
```
df['Release_Date'].head()
```

Out[79]:
```
0    2021
1    2022
2    2022
3    2021
4    2021
Name: Release_Date, dtype: int32
```

In [81]:
```
df.head()
```

Out[81]:

	Release_Date	Title	Overview	Popularity	Vote_Count	Vote_Average	Original_Lan
0	2021	Spider-Man: No Way Home	Peter Parker is unmasked and no longer able to...	5083.954	8940	8.3	
1	2022	The Batman	In his second year of fighting crime, Batman u...	3827.658	1151	8.1	
2	2022	No Exit	Stranded at a rest stop in the mountains durin...	2618.087	122	6.3	
3	2021	Encanto	The tale of an extraordinary family, the Madri...	2402.201	5076	7.7	
4	2021	The King's Man	As a collection of history's worst tyrants and...	1895.511	1793	7.0	

Dropping **Overview**, **Original_Languege** and **Poster_Url**

In [84]:
```
col = ['Overview','Original_Language','Poster_Url']

df.drop(col , axis=1 , inplace = True)
```

In [86]:
```
df.columns
```

Out[86]:
```
Index(['Release_Date', 'Title', 'Popularity', 'Vote_Count', 'Vote_Average',
       'Genre'],
      dtype='object')
```

In [88]:
```
df.head()
```

Out[88]:

	Release_Date	Title	Popularity	Vote_Count	Vote_Average	Genre
0	2021	Spider-Man: No Way Home	5083.954	8940	8.3	Action, Adventure, Science Fiction
1	2022	The Batman	3827.658	1151	8.1	Crime, Mystery, Thriller
2	2022	No Exit	2618.087	122	6.3	Thriller
3	2021	Encanto	2402.201	5076	7.7	Animation, Comedy, Family, Fantasy
4	2021	The King's Man	1895.511	1793	7.0	Action, Adventure, Thriller, War

categorizing Vote_Average column

We would cut the **Vote_Average** values and make 4 categories: **popular**, **average**, **below_avg**, **not_popular** to describe it more using **catigorize_col()** function provided above.

In [91]:
```
def catigorize_col(df , col , labels):

    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
```

In [93]:
```
labels = ['not_popular', 'below_avg', 'average', 'popular']

catigorize_col(df , 'Vote_Average' , labels)
df['Vote_Average'].unique()
```

Out[93]:
```
['popular', 'below_avg', 'average', 'not_popular', NaN]
Categories (4, object): ['not_popular' < 'below_avg' < 'average' < 'popular']
```

In [96]:
```
df.head()
```

Out[96]:

	Release_Date	Title	Popularity	Vote_Count	Vote_Average	Genre
0	2021	Spider-Man: No Way Home	5083.954	8940	popular	Action, Adventure, Science Fiction
1	2022	The Batman	3827.658	1151	popular	Crime, Mystery, Thriller
2	2022	No Exit	2618.087	122	below_avg	Thriller
3	2021	Encanto	2402.201	5076	popular	Animation, Comedy, Family, Fantasy
4	2021	The King's Man	1895.511	1793	average	Action, Adventure, Thriller, War

In [98]:
```
df['Vote_Average'].value_counts()
```

Out[98]:
```
Vote_Average
not_popular    2467
popular        2450
average        2412
below_avg      2398
Name: count, dtype: int64
```

In [104…
```
df.dropna(inplace = True)

df.isna().sum()
```

Out[104…
```
Release_Date    0
Title           0
Popularity      0
Vote_Count      0
Vote_Average    0
Genre           0
dtype: int64
```

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

In [111…
```
df['Genre'] = df['Genre'].str.split(", ")

df['Genre']
```

Out[111…
```
0       [Action, Adventure, Science Fiction]
1                 [Crime, Mystery, Thriller]
2                                  [Thriller]
3       [Animation, Comedy, Family, Fantasy]
4         [Action, Adventure, Thriller, War]
                        ...
9822                          [Drama, Crime]
9823                                [Horror]
9824             [Mystery, Thriller, Horror]
9825                 [Music, Drama, History]
9826           [War, Drama, Science Fiction]
Name: Genre, Length: 9727, dtype: object
```

In [113…
```
df = df.explode('Genre').reset_index(drop=True)
df.head()
```

Out[113…

	Release_Date	Title	Popularity	Vote_Count	Vote_Average	Genre
0	2021	Spider-Man: No Way Home	5083.954	8940	popular	Action
1	2021	Spider-Man: No Way Home	5083.954	8940	popular	Adventure
2	2021	Spider-Man: No Way Home	5083.954	8940	popular	Science Fiction
3	2022	The Batman	3827.658	1151	popular	Crime
4	2022	The Batman	3827.658	1151	popular	Mystery

In [115…
```
df['Genre'] = df['Genre'].astype("category")

print(df['Genre'].dtype)
```

```
category
```

In [117…
```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 25552 entries, 0 to 25551
Data columns (total 6 columns):
 #   Column        Non-Null Count  Dtype   
---  ------        --------------  -----   
 0   Release_Date  25552 non-null  int32   
 1   Title         25552 non-null  object  
 2   Popularity    25552 non-null  float64 
 3   Vote_Count    25552 non-null  int64   
 4   Vote_Average  25552 non-null  category
 5   Genre         25552 non-null  category
dtypes: category(2), float64(1), int32(1), int64(1), object(1)
memory usage: 749.6+ KB
```

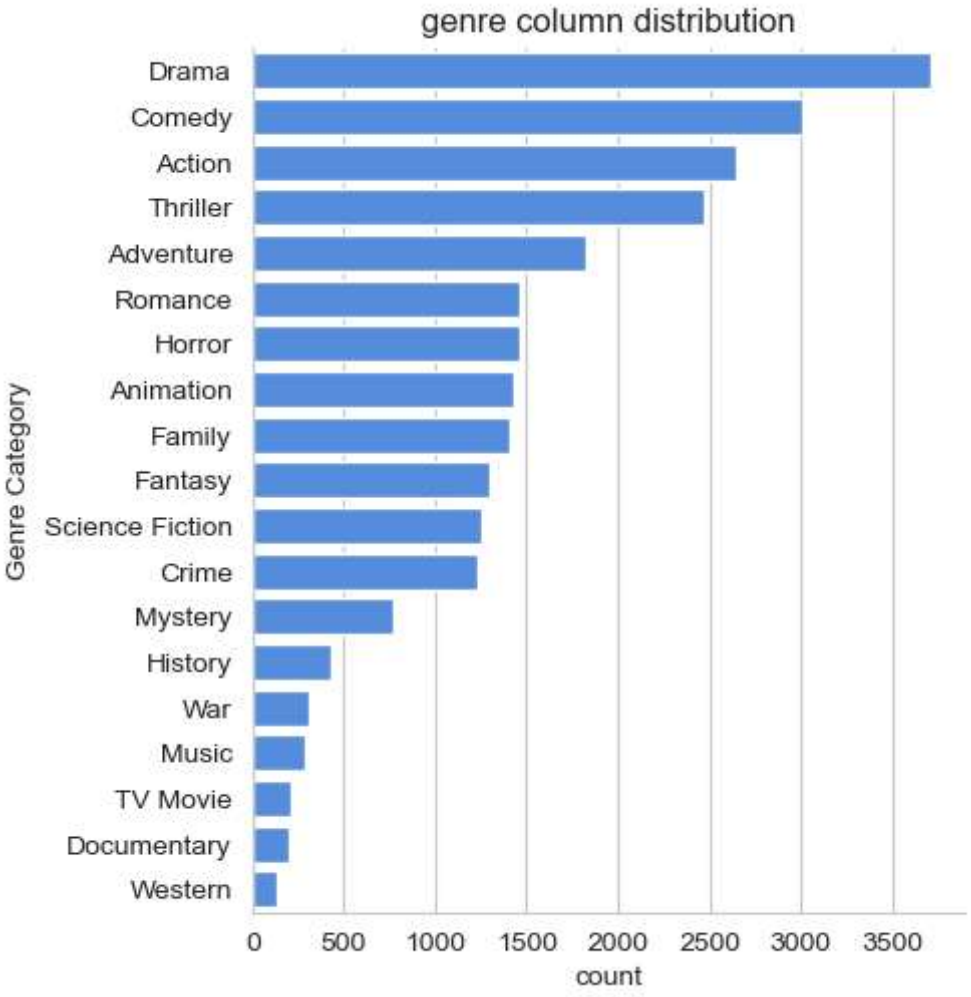
Data Visualization

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

In [121…
```
sns.set_style('whitegrid')
```

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

In [138…
```
sns.catplot(
    y = 'Genre',
    data = df,
    kind = "count",
    order = df['Genre'].value_counts().index,
    color = '#4287f5'
)
plt.title("genre column distribution")
plt.ylabel("Genre Category")
plt.show()
```



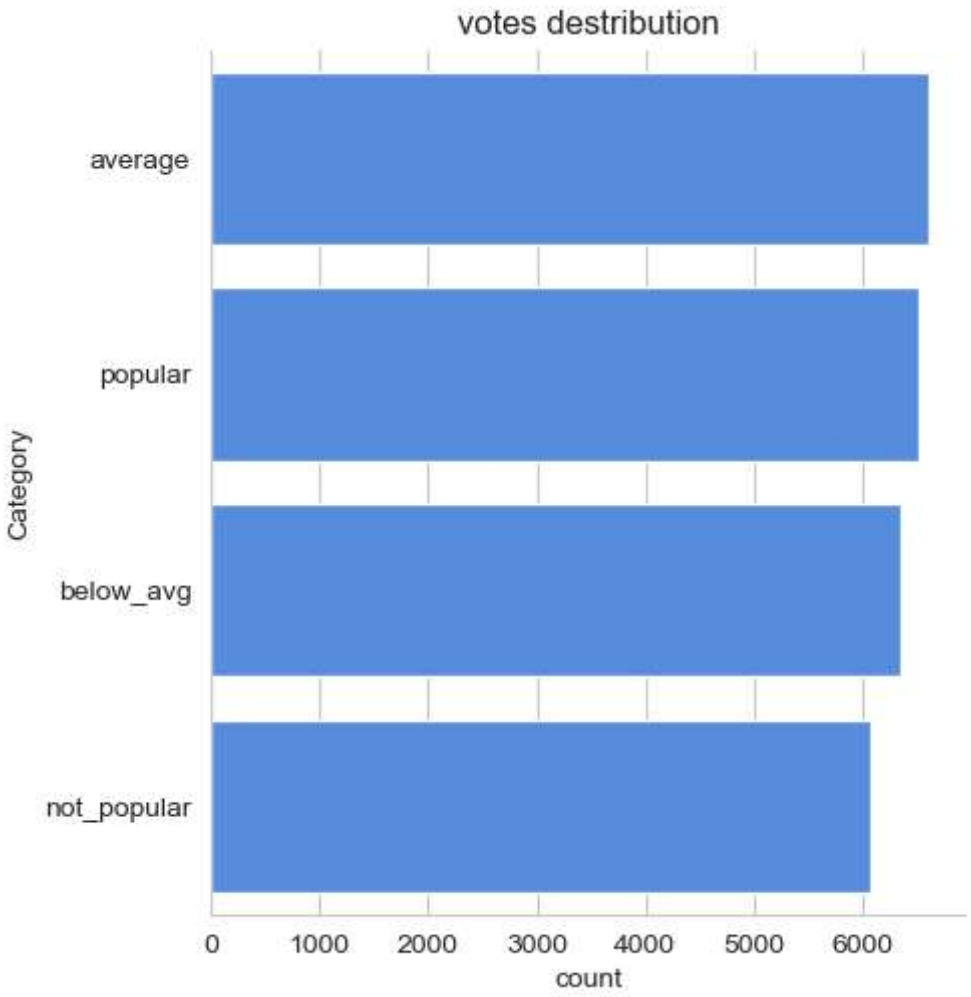
- we can notice from the above visual that Drama genre is the most frequent genre

in our dataset and has appeared more than 14% of the times among 19 other genres.

Q2: What genres has highest votes ?

In [146…

```
sns.catplot(
    y = 'Vote_Average',
    data = df,
    kind = "count",
    order = df['Vote_Average'].value_counts().index,
    color = '#4287f5'
)
plt.title("votes destribution")
plt.ylabel("Category")
plt.show()
```



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

In [151…

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

Out[151…

	Release_Date	Title	Popularity	Vote_Count	Vote_Average	Genre
0	2021	Spider-Man: No Way Home	5083.954	8940	popular	Action
1	2021	Spider-Man: No Way Home	5083.954	8940	popular	Adventure
2	2021	Spider-Man: No Way Home	5083.954	8940	popular	Science Fiction

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

In [154…

```
df[df['Popularity'] == df['Popularity'].min()]
```

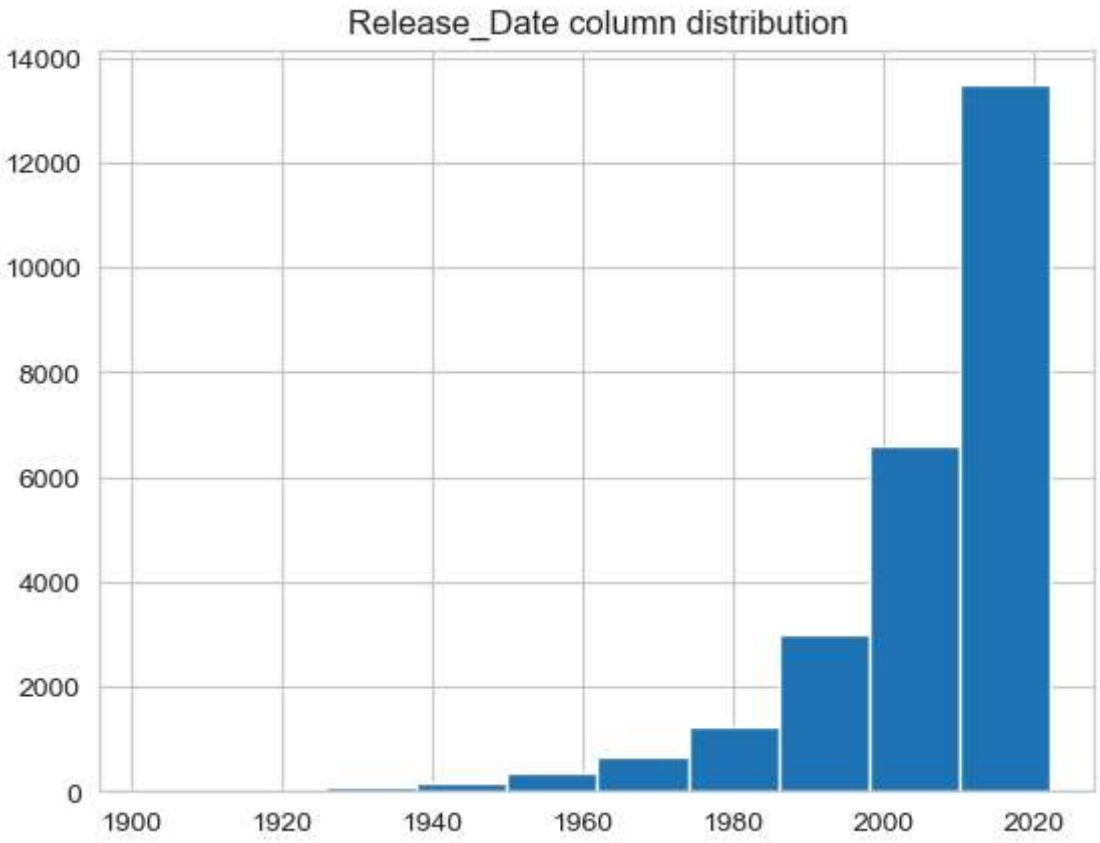
Out[154…

	Release_Date	Title	Popularity	Vote_Count	Vote_Average	Genre
25546	2021	The United States vs. Billie Holiday	13.354	152	average	Music
25547	2021	The United States vs. Billie Holiday	13.354	152	average	Drama
25548	2021	The United States vs. Billie Holiday	13.354	152	average	History
25549	1984	Threads	13.354	186	popular	War
25550	1984	Threads	13.354	186	popular	Drama
25551	1984	Threads	13.354	186	popular	Science Fiction

Q5: Which year has the most filmmed movies?

In [175…

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



In [171…
```
df.head()
```

Out[171…

	Release_Date	Title	Popularity	Vote_Count	Vote_Average	Genre
0	2021	Spider-Man: No Way Home	5083.954	8940	popular	Action
1	2021	Spider-Man: No Way Home	5083.954	8940	popular	Adventure
2	2021	Spider-Man: No Way Home	5083.954	8940	popular	Science Fiction
3	2022	The Batman	3827.658	1151	popular	Crime
4	2022	The Batman	3827.658	1151	popular	Mystery

Conclusion

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

- 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?

- 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?

- 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?

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

Q5: Which year has the most filmmed movies?

- year 2020 has the highest filmming rate in our dataset.

In [ ]: