- Import libraries

In [48]:

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

- Get Data From CSV

In [49]:

```
df=pd.read_csv('mymovie.csv',lineterminator='\n')
```

- View First 5 Rows of DataFrame

In [50]:

```
df.head()
```

Out[50]:

	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	

- Checking DataFrame Info (Column Types & Nulls)

```
In [51]: 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
```

- Viewing First 5 Values of 'Genre' Column

```
In [52]: df['Genre'].head()
```

```
Out[52]: 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
```

- Counting Duplicate Rows in DataFrame

```
In [53]: df.duplicated().sum()
```

```
Out[53]: np.int64(0)
```

- Describe DataFrame Stats

```
In [54]: df.describe()
```

Out[54]:

	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

- we have a dataframe consisting of 9827 rows 9 columns.
- our dataset looks a bit tidy with no NaNs or duplicated Values.
- Release_Date column needs to be casted into date time and to extract only the year value.
- Overview, Original_Language and Poster-url wouldn't be so useful during analysis, so we'll drop them.
- there is noticable outliers in Popularity column.
- Vote_Average better be categorised for proper analysis.
- Genre column has comma sepereted values and white spaces that needs to be handled and casted into category. Exploration Summary.

- Converting 'Release_Date' Column to Date Format

In [55]:

```
df['Release_Date']=pd.to_datetime(df['Release_Date'])
print(df['Release_Date'].dtypes)
```

```
datetime64[ns]
```

- Extracting Year from 'Release_Date

In [56]:

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

Out[56]:

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

- Dropping Unwanted Columns from DataFrame

In [57]:

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

In [58]:

```
df.head()
```

Out[58]:

	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 and make 4 categories -popular, average, below_avg, not_popular to describge it more using catigorize_col() funciton provided above.

In [59]:

```
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 [60]:

```
labels=['not_popular','below_avg','average','popular']
catigorize_col(df,'Vote_Average',labels)
df['Vote_Average'].unique()
```

Out[60]:

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

In [61]:

```
df.head()
```

Out[61]:

	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

- Counting Frequency of Each Vote_Average Value

In [62]:

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

Out[62]:

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

- Removing Missing Values and Checking Again

In [63]:

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

Out[63]:

```
Release_Date    0
Title           0
Popularity      0
Vote_Count      0
Vote_Average    0
Genre           0
dtype: int64
```

In [64]:

```
df.head()
```

Out[64]:

	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

- Splitting and Expanding Genre Column into Multiple Rows

In [65]:

```
df['Genre']=df['Genre'].str.split(', ')
df=df.explode('Genre').reset_index(drop=True)
df.head()
```

Out[65]:

	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

- Converting Genre Column to Categorical Type

In [66]:

```
df['Genre']=df['Genre'].astype('category')
df['Genre'].dtype
```

Out[66]:

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

- Counting Unique Values in Each Column

In [67]:

```
df.nunique()
```

Out[67]:

```
Release_Date     100
Title           9415
Popularity      8088
Vote_Count      3265
Vote_Average       4
Genre             19
dtype: int64
```

In [68]:

```
df.head()
```

Out[68]:

	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

- Data Visualization

- Apply Whitegrid Style to Charts

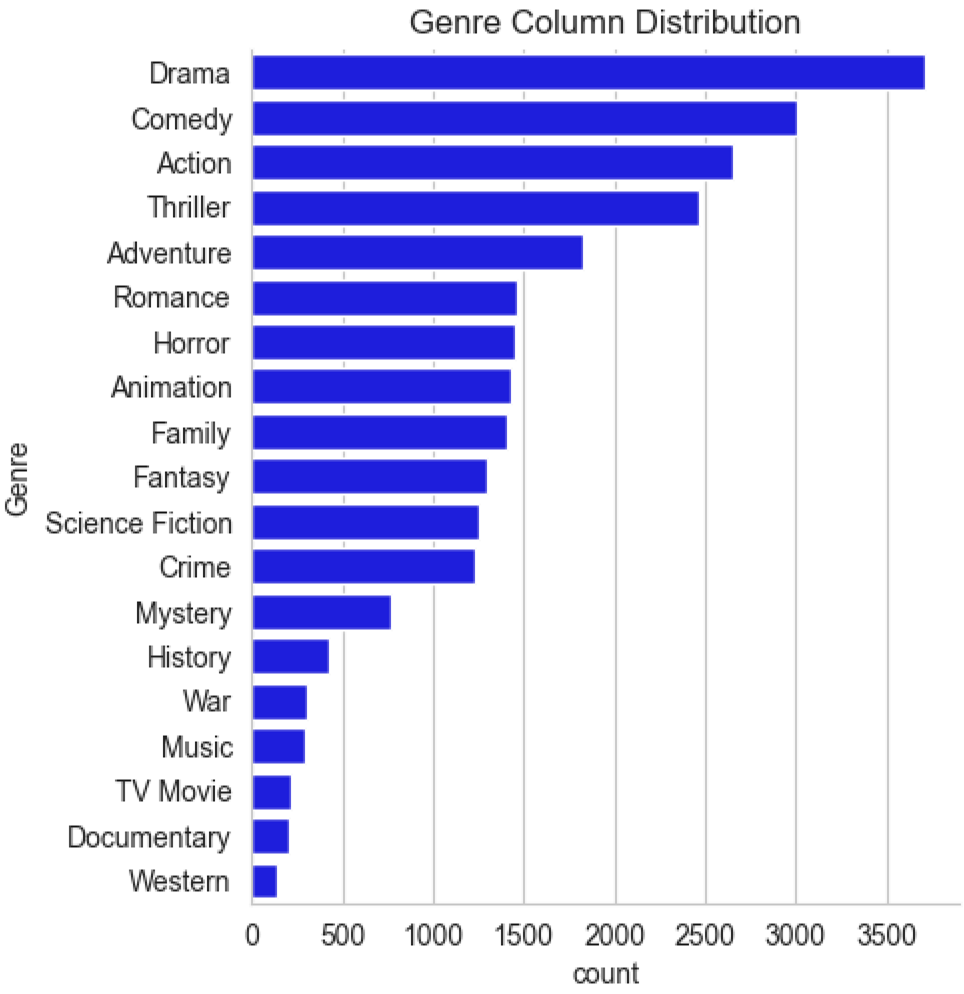
In [69]:

```
sns.set_style('whitegrid')
```

- Plotting Genre Distribution Using Seaborn

In [83]:

```
sns.catplot(y='Genre',data=df,kind='count',order=df['Genre'].value_counts().index,c
plt.title('Genre Column Distribution')
plt.show()
```



In [84]:

```
df.head()
```

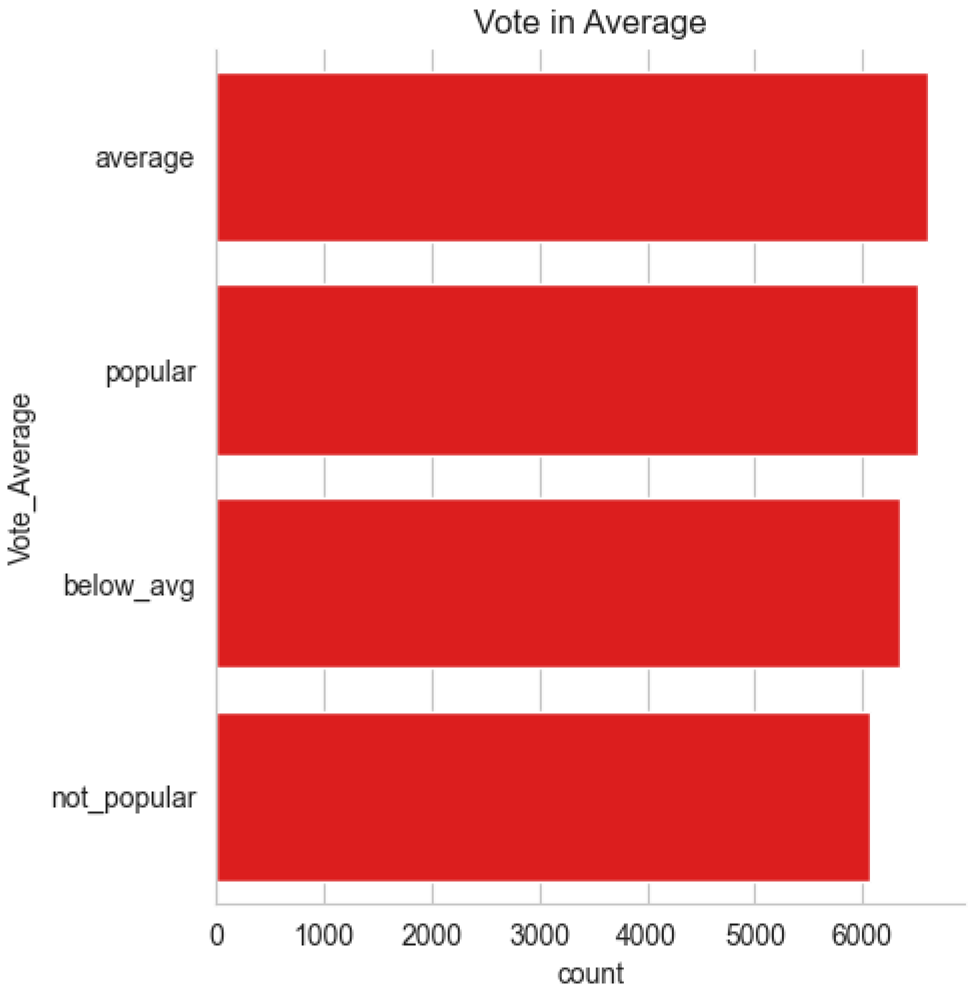
Out[84]:

	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

- Plotting Vote_Average Distribution Using Seaborn

In [89]:

```
sns.catplot(y='Vote_Average',data=df,kind='count',order=df['Vote_Average'].value_co
plt.title('Vote in Average')
plt.show()
```



- Finding the Most Popular Movie

```
In [90]: df[df['Popularity']==df['Popularity'].max()]
```

Out[90]:

	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

- Finding the Least Popular Movie

```
In [92]: df[df['Popularity']==df['Popularity'].min()]
```

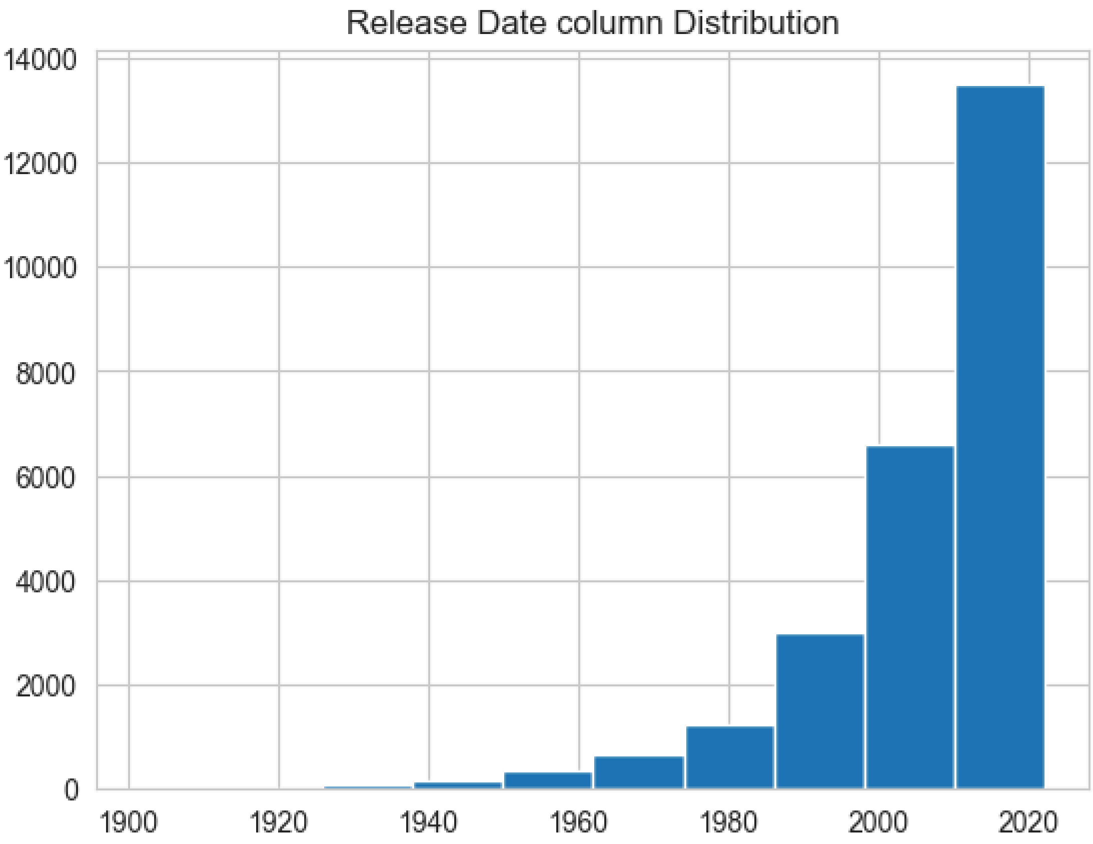
Out[92]:

	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

- Histogram of Release Year Distribution

In [93]:

```
df['Release_Date'].hist()
plt.title('Release Date column Distribution')
plt.show()
```



- 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 Action , genre ?
- Spider-Man: No Way Home has the highest popularity rate in our dataset and it has genres of Adventure and Sience Fiction .
- Q3: 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`.
- Q4: Which year has the most filmmed movies?
- year 2020 has the highest filmming rate in our dataset