In [1]:
```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns

# Pandas display settings
pd.set_option("display.max_columns", None)
pd.set_option("display.max_colwidth", None)

df = pd.read_csv("mymoviedb.csv", lineterminator="\n")
df.head()
```

Out[1]:

	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 separate his normal life from the high-stakes of being a super-hero. When he asks for help from Doctor Strange the stakes become even more dangerous, forcing him to discover what it truly means to be Spider-Man.	5083.954	8940	8.3	
1	2022-03-01	The Batman	In his second year of fighting crime, Batman uncovers corruption in Gotham City that connects to his own family while facing a serial killer known as the Riddler.	3827.658	1151	8.1	
2	2022-02-25	No Exit	Stranded at a rest stop in the mountains during a blizzard, a recovering addict discovers a kidnapped child hidden in a car belonging to	2618.087	122	6.3	

	Release_Date	Title	Overview	Popularity	Vote_Count	Vote_Average	Original_Lan
			one of the people inside the building which sets her on a terrifying struggle to identify who among them is the kidnapper.				
3	2021-11-24	Encanto	The tale of an extraordinary family, the Madrigals, who live hidden in the mountains of Colombia, in a magical house, in a vibrant town, in a wondrous, charmed place called an Encanto. The magic of the Encanto has blessed every child in the family with a unique gift from super strength to the power to heal—every child except one, Mirabel. But when she discovers that the magic surrounding the Encanto is in danger, Mirabel decides that	2402.201	5076	7.7	

	Release_Date	Title	Overview	Popularity	Vote_Count	Vote_Average	Original_Lan
			she, the only ordinary Madrigal, might just be her exceptional family's last hope.				
4	2021-12-22	The King's Man	As a collection of history's worst tyrants and criminal masterminds gather to plot a war to wipe out millions, one man must race against time to stop them.	1895.511	1793	7.0	

In [2]:
```
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
```

In [3]:
```
df['Genre'].head()
```

Out[3]:
```
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
```

In [4]:
```
df.duplicated().sum()
```

Out[4]:
```
0
```

In [5]:
```
df.describe()
```

Out[5]:

	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

In [6]:
```
df.head()
```

Out[6]:

	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 separate his normal life from the high-stakes of being a super-hero. When he asks for help from Doctor Strange the stakes become even more dangerous, forcing him to discover what it truly means to be Spider-Man.	5083.954	8940	8.3	
1	2022-03-01	The Batman	In his second year of fighting crime, Batman uncovers corruption in Gotham City that connects to his own family while facing a serial killer known as the Riddler.	3827.658	1151	8.1	
2	2022-02-25	No Exit	Stranded at a rest stop in the mountains during a blizzard, a recovering addict discovers a kidnapped child hidden in a car belonging to	2618.087	122	6.3	

	Release_Date	Title	Overview	Popularity	Vote_Count	Vote_Average	Original_Lan
			one of the people inside the building which sets her on a terrifying struggle to identify who among them is the kidnapper.				
3	2021-11-24	Encanto	The tale of an extraordinary family, the Madrigals, who live hidden in the mountains of Colombia, in a magical house, in a vibrant town, in a wondrous, charmed place called an Encanto. The magic of the Encanto has blessed every child in the family with a unique gift from super strength to the power to heal—every child except one, Mirabel. But when she discovers that the magic surrounding the Encanto is in danger, Mirabel decides that	2402.201	5076	7.7	

	Release_Date	Title	Overview	Popularity	Vote_Count	Vote_Average	Original_Lan
			she, the only ordinary Madrigal, might just be her exceptional family's last hope.				
4	2021-12-22	The King's Man	As a collection of history's worst tyrants and criminal masterminds gather to plot a war to wipe out millions, one man must race against time to stop them.	1895.511	1793	7.0	

In [7]:
```
df['Release_Date']=pd.to_datetime(df['Release_Date'])
print(df['Release_Date'].dtypes)
```

```
datetime64[ns]
```

In [8]:
```
df['Release_Date']= df['Release_Date'].dt.year
df['Release_Date'].dtypes
```

Out[8]:
```
dtype('int32')
```

In [9]:
```
df.head()
```

Out[9]:

	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 separate his normal life from the high-stakes of being a super-hero. When he asks for help from Doctor Strange the stakes become even more dangerous, forcing him to discover what it truly means to be Spider-Man.	5083.954	8940	8.3	
1	2022	The Batman	In his second year of fighting crime, Batman uncovers corruption in Gotham City that connects to his own family while facing a serial killer known as the Riddler.	3827.658	1151	8.1	
2	2022	No Exit	Stranded at a rest stop in the mountains during a blizzard, a recovering addict discovers a kidnapped child hidden in a car belonging to	2618.087	122	6.3	

	Release_Date	Title	Overview	Popularity	Vote_Count	Vote_Average	Original_Lan
			one of the people inside the building which sets her on a terrifying struggle to identify who among them is the kidnapper.				
3	2021	Encanto	The tale of an extraordinary family, the Madrigals, who live hidden in the mountains of Colombia, in a magical house, in a vibrant town, in a wondrous, charmed place called an Encanto. The magic of the Encanto has blessed every child in the family with a unique gift from super strength to the power to heal—every child except one, Mirabel. But when she discovers that the magic surrounding the Encanto is in danger, Mirabel decides that	2402.201	5076	7.7	

	Release_Date	Title	Overview	Popularity	Vote_Count	Vote_Average	Original_Lan
			she, the only ordinary Madrigal, might just be her exceptional family's last hope.				
4	2021	The King's Man	As a collection of history's worst tyrants and criminal masterminds gather to plot a war to wipe out millions, one man must race against time to stop them.	1895.511	1793	7.0	

Dropping the columns

In [10]:
```
cols = ['Overview','Original_Language','Poster_Url']
df.drop(cols, axis=1, inplace=True)
df.columns
```

Out[10]:
```
Index(['Release_Date', 'Title', 'Popularity', 'Vote_Count', 'Vote_Average',
       'Genre'],
      dtype='object')
```

In [11]:
```
df.head()
```

Out[11]:

	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

In [12]:
```
def categorize_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 [13]:
```
labels = ['not_popular','below_avg' , 'average','popular']

categorize_col(df,'Vote_Average',labels)

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

Out[13]:
```
['popular', 'below_avg', 'average', 'not_popular', NaN]
Categories (4, object): ['not_popular' < 'below_avg' < 'average' < 'popular']
```

In [14]:
```
df.head()
```

Out[14]:

	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 [15]:
```
df['Vote_Average'].value_counts()
```

Out[15]:
```
Vote_Average
not_popular    2467
popular        2450
average        2412
below_avg      2398
Name: count, dtype: int64
```

In [16]:
```
df.dropna(inplace = True)

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

Out[16]:
```
Release_Date    0
Title           0
Popularity      0
Vote_Count      0
Vote_Average    0
Genre           0
dtype: int64
```

In [17]:
```
df.head()
```

Out[17]:

	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

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

In [18]:
```
df['Genre']= df['Genre'].str.split(', ')

df = df.explode('Genre').reset_index(drop = True)
df.head()
```

Out[18]:

	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 [19]:
```
#casting column into category

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

df['Genre'].dtypes
```

Out[19]:
```
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)
```

In [20]:
```
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
```

In [21]:
```
df.nunique()
```

Out[21]:
```
Release_Date     100
Title           9415
Popularity      8088
Vote_Count      3265
Vote_Average       4
Genre             19
dtype: int64
```

In [22]:
```
df.head()
```

Out[22]:

	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

In [23]:
```
sns.set_style('whitegrid')
```

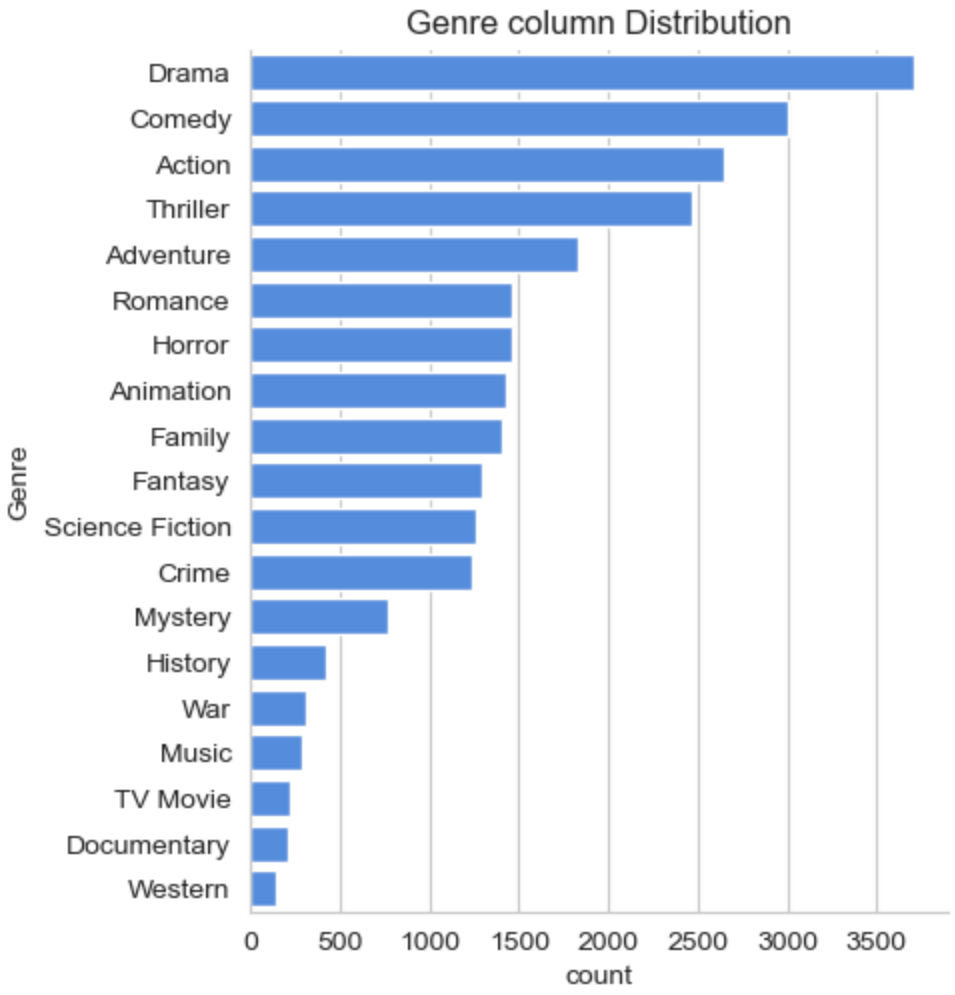
What is the most frequent genre of movies released on Netflix ?

In [24]:
```
df['Genre'].describe()
```

Out[24]:
```
count     25552
unique       19
top       Drama
freq       3715
Name: Genre, dtype: object
```

In [25]:
```
sns.catplot(y = 'Genre',data = df , kind= 'count', order= df['Genre'].value_counts(
plt.title ('Genre column Distribution')
plt.show()
```

```
c:\Users\user\anaconda3\Lib\site-packages\seaborn\categorical.py:641: FutureWarning:
The default of observed=False is deprecated and will be changed to True in a future
version of pandas. Pass observed=False to retain current behavior or observed=True t
o adopt the future default and silence this warning.
  grouped_vals = vals.groupby(grouper)
c:\Users\user\anaconda3\Lib\site-packages\seaborn\categorical.py:641: FutureWarning:
The default of observed=False is deprecated and will be changed to True in a future
version of pandas. Pass observed=False to retain current behavior or observed=True t
o adopt the future default and silence this warning.
  grouped_vals = vals.groupby(grouper)
```



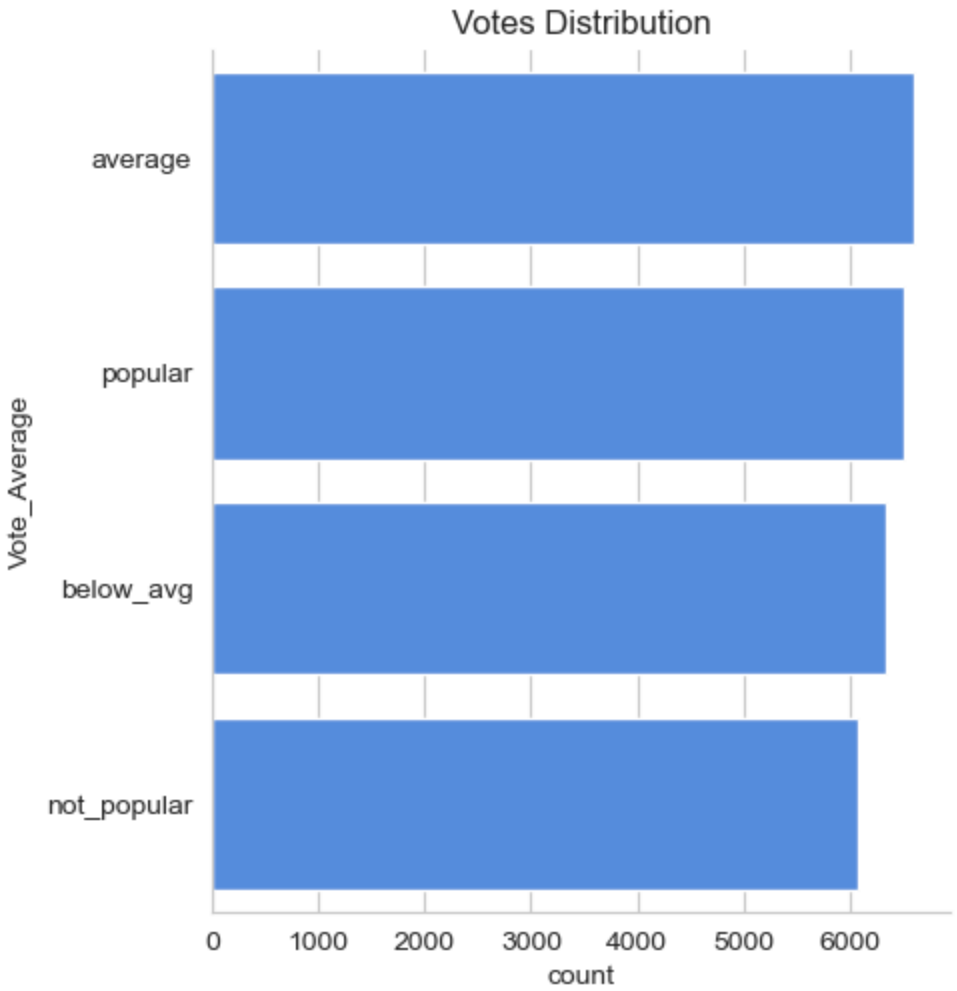
Which has highest votes in vote avg column ?

In [26]:
```
sns.catplot(y = 'Vote_Average', data = df , kind= 'count',
            order= df['Vote_Average'].value_counts().index, color= '#4287f5')
```

```
plt.title('Votes Distribution')

plt.show()
```

```
c:\Users\user\anaconda3\Lib\site-packages\seaborn\categorical.py:641: FutureWarning:
The default of observed=False is deprecated and will be changed to True in a future
version of pandas. Pass observed=False to retain current behavior or observed=True t
o adopt the future default and silence this warning.
  grouped_vals = vals.groupby(grouper)
c:\Users\user\anaconda3\Lib\site-packages\seaborn\categorical.py:641: FutureWarning:
The default of observed=False is deprecated and will be changed to True in a future
version of pandas. Pass observed=False to retain current behavior or observed=True t
o adopt the future default and silence this warning.
  grouped_vals = vals.groupby(grouper)
```



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

In [27]:
```
df.head(2)
```

Out[27]:

	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

In [28]:
```
df[df['Popularity'] == df['Popularity'].max()]
```

Out[28]:

	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

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

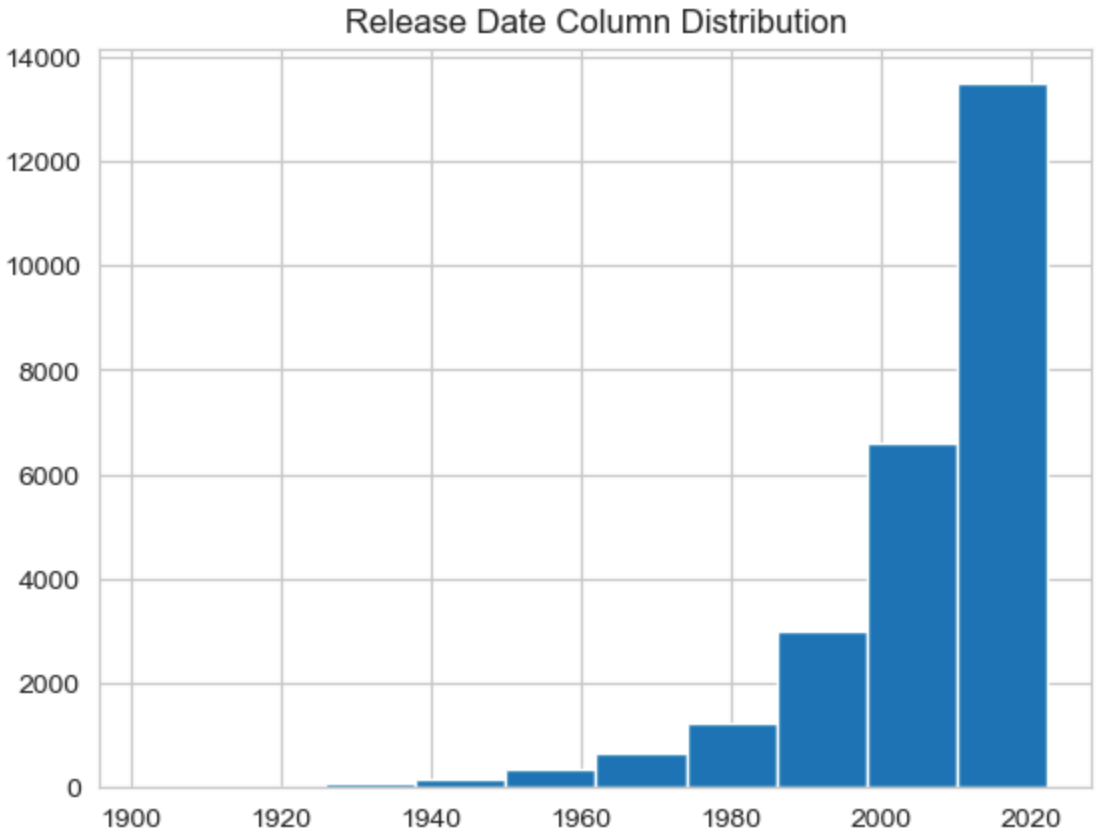
In [29]:
```
df[df['Popularity'] == df['Popularity'].min()]
```

Out[29]:

	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

Which year has the most filmmed movies?

In [30]:
```
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 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.