

```
In [13]: import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns

In [14]: df=pd.read_csv('mymoviedb.csv',lineterminator='\n')
df.head()
```

Out [14]:	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.tmbd.org/t/p/original/1g0dhYtq4l...
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.tmbd.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.tmbd.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.tmbd.org/t/p/original/4jDPNhkMr5...
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.tmbd.org/t/p/original/aq4Pww5Xeu...

```
In [12]: 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 [5]: df['Genre'].head()
```

```
Out [5]: 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 [7]: df.duplicated().sum()

Out [7]: 0

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

Out [9]:	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 [16]: df['Release_Date'] = pd.to_datetime(df['Release_Date'])
# confirming changes
print(df['Release_Date'].dtypes)

datetime64[ns]

In [18]: df['Release_Date'] = df['Release_Date'].dt.year
df['Release_Date'].dtypes
```

```
Out [18]: dtype('int32')

Dropping the columns
```

```
In [24]: cols=['Overview','Original_Language','Poster_Url']
df.drop(cols,axis=1,inplace=True)
df.columns
```

```
Out [24]: Index(['Release_Date', 'Title', 'Popularity', 'Vote_Count', 'Vote_Average',
              'Genre'],
              dtype='object')
```

```
In [26]: df.head()
```

Out [26]:	Release_Date	Title	Popularity	Vote_Count	Vote_Average	Genre
0	2021-12-15	Spider-Man: No Way Home	5083.954	8940	8.3	Action, Adventure, Science Fiction
1	2022-03-01	The Batman	3827.658	1151	8.1	Crime, Mystery, Thriller
2	2022-02-25	No Exit	2618.087	122	6.3	Thriller
3	2021-11-24	Encanto	2402.201	5076	7.7	Animation, Comedy, Family, Fantasy
4	2021-12-22	The King's Man	1895.511	1793	7.0	Action, Adventure, Thriller, War

```
In [35]: df['Vote_Average'].unique()

Out [35]: ['popular', 'below_avg', 'average', 'not_popular', NaN]
Categories (4, object): ['not_popular' < 'below_avg' < 'average' < 'popular']

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

```
In [38]: df.dropna(inplace=True)

In [39]: df.isna().sum()

Out [39]: Release_Date    0
Title                  0
Popularity             0
Vote_Count             0
Vote_Average          0
Genre                 0
dtype: int64
```

```
In [40]: df['Genre']=df['Genre'].str.split(',')
df=df.explode('Genre').reset_index(drop=True)
df.head()
```

```
Out [40]:
```

	Release_Date	Title	Popularity	Vote_Count	Vote_Average	Genre
0	2021-12-15	Spider-Man: No Way Home	5083.954	8940	8.3	Action
1	2021-12-15	Spider-Man: No Way Home	5083.954	8940	8.3	Adventure
2	2021-12-15	Spider-Man: No Way Home	5083.954	8940	8.3	Science Fiction
3	2022-03-01	The Batman	3827.658	1151	8.1	Crime
4	2022-03-01	The Batman	3827.658	1151	8.1	Mystery

Casting columns into Categories

```
In [44]: df['Genre']=df['Genre'].astype('category')
df['Genre'].dtypes
```

```
Out [44]: 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)
```

```
In [45]: df.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 25793 entries, 0 to 25792
Data columns (total 6 columns):
 #   Column                Non-Null Count  Dtype  
---  --
 0   Release_Date          25793 non-null object  
 1   Title                 25793 non-null object  
 2   Popularity            25793 non-null float64  
 3   Vote_Count            25793 non-null int64  
 4   Vote_Average          25793 non-null float64  
 5   Genre                 25793 non-null category
dtypes: category(1), float64(2), int64(1), object(2)
memory usage: 1.0+ MB

In [46]: df.nunique()
```

```
Out [46]: Release_Date    5893
Title                9513
Popularity           8160
Vote_Count           3266
Vote_Average         74
Genre                 19
dtype: int64
```

Data Visualization

```
In [49]: sns.set_style('whitegrid')
```

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

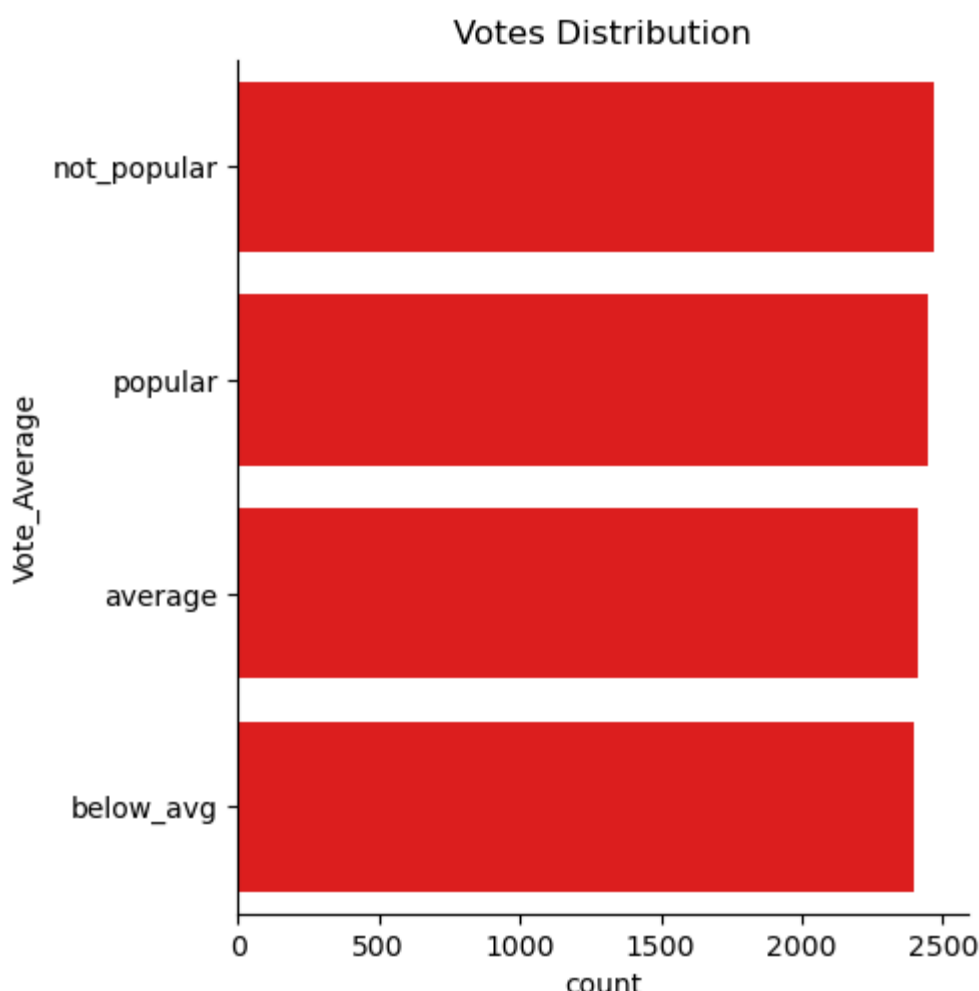
```
In [51]: df['Genre'].describe()

Out [51]: count      25793
unique         19
top      Drama
freq         3744
Name: Genre, dtype: object

In [ ]: # 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()
```

Which is the highest vote in the vote column?

```
In [27]: sns.catplot(y='Vote_Average',data=df,kind='count',
order=df['Vote_Average'].value_counts().index,
colours='red')
plt.title('Votes Distribution')
plt.show()
```



What movie got the highest popularity? What is its genre?

```
In [61]: df[df['Popularity']==df['Popularity'].max()]

Out [61]:
```

	Release_Date	Title	Popularity	Vote_Count	Vote_Average	Genre
0	2021-12-15	Spider-Man: No Way Home	5083.954	8940	8.3	Action
1	2021-12-15	Spider-Man: No Way Home	5083.954	8940	8.3	Adventure
2	2021-12-15	Spider-Man: No Way Home	5083.954	8940	8.3	Science Fiction

What movie got the low popularity? What is its genre?

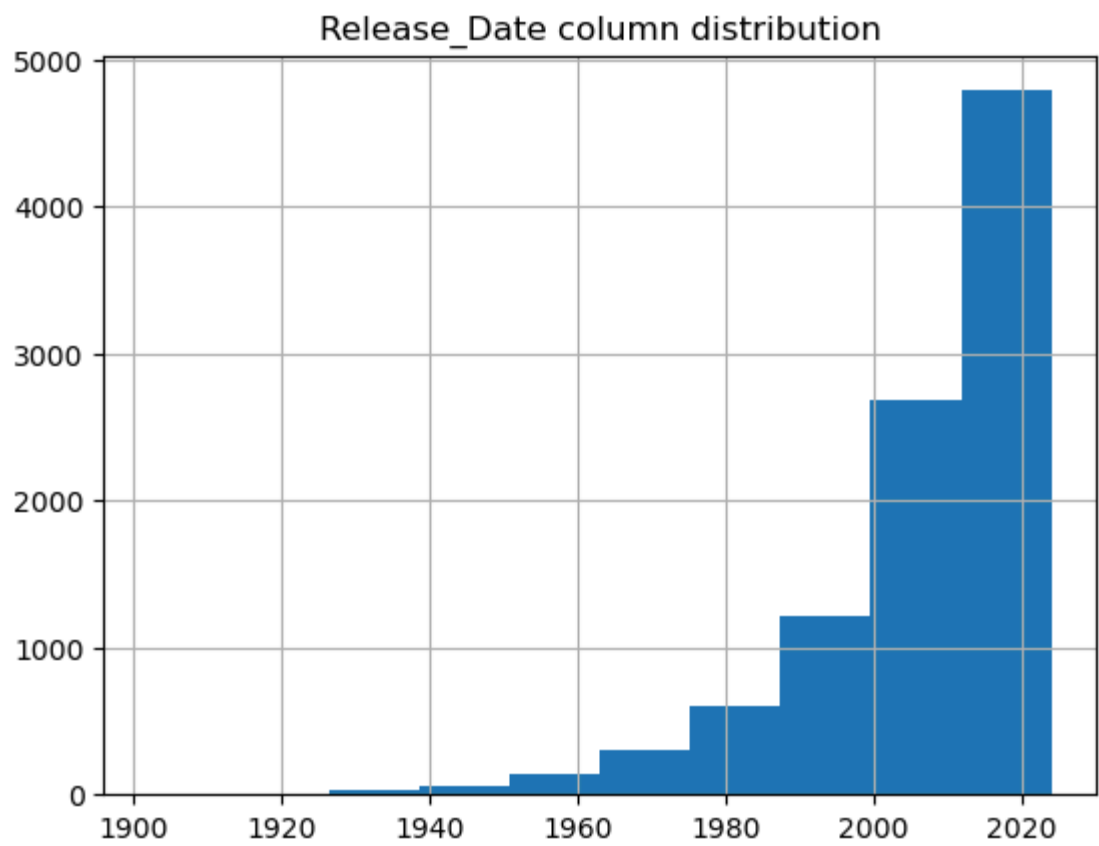
```
In [65]: df[df['Popularity']==df['Popularity'].min()]
df.head(2)

Out [65]:
```

	Release_Date	Title	Popularity	Vote_Count	Vote_Average	Genre
0	2021-12-15	Spider-Man: No Way Home	5083.954	8940	8.3	Action
1	2021-12-15	Spider-Man: No Way Home	5083.954	8940	8.3	Adventure

which year of the most filmed movies

```
In [26]: df['Release_Date'].hist()
plt.title('Release_Date column distribution')
plt.show()
```



```
In [ ]:

In [ ]:
```