Netflix data analysis

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

In [2]:
```
df = pd.read_csv('netflix_titles.csv',lineterminator '\n')
```

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

Out[3]:

	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/1g0dhYtq4...
1	2022-03-01	The Batman	In his second year of fighting crime, Batman u...	3627.658	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/4EP9nN4MA5...
4	2021-12-22	The King's Man	As a collection of history's worst tyrants and...	1895.511	1750	7.0	en	Action, Adventure, Thriller, War	https://image.tmdb.org/t/p/original/aq4Pwv5Xeu...

In [4]:
```
df.info()
```

Out[4]:
```
<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
dtype: float64(2), int64(2), object(5)
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 [6]:
```
df.duplicated().sum()
```

Out[6]:
```
np.int64(0)
```

In [7]:
```
df.describe()
```

Out[7]:

	Popularity	Vote_Count	Vote_Average
count	9827.000000	9827.000000	9827.000000
mean	40.326088	1392.805536	6.439534
std	108.873968	2611.206907	1.129759
min	13.354000	0.000000	0.500000
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

1-there have dataframe consisting of 9827 rows and 9 columns. 2-#no duplicates. 3-#release_date column needs to be casted into date time and to extract only the year value. 4_#overview,original_language and poster-url not so useful,so we'l drop the. 5_#outlier in popularity column. 6_#vote_average better be categorised for proper analysis. 7_#genre column has cone seprate value and white spaces that needs to be handled and casted into category.

In [8]:
```
df.head()
```

Out[8]:

	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/1g0dhYtq4...
1	2022-03-01	The Batman	In his second year of fighting crime, Batman u...	3627.658	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/4EP9nN4MA5...
4	2021-12-22	The King's Man	As a collection of history's worst tyrants and...	1895.511	1750	7.0	en	Action, Adventure, Thriller, War	https://image.tmdb.org/t/p/original/aq4Pwv5Xeu...

In [9]:
```
df['Release_Date'] = pd.to_datetime(df['Release_Date'])

print(df['Release_Date'].dtypes)

datetime64[ns]
```

In [10]:
```
df['Release_Date']=df['Release_Date'].dt.year

df['Release_Date'].dtypes
```

Out[10]:
```
dtype('int32')
```

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

Out[11]:

	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/1g0dhYtq4...
1	2022	The Batman	In his second year of fighting crime, Batman u...	3627.658	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/4EP9nN4MA5...
4	2021	The King's Man	As a collection of history's worst tyrants and...	1895.511	1750	7.0	en	Action, Adventure, Thriller, War	https://image.tmdb.org/t/p/original/aq4Pwv5Xeu...

dropping the columns

In [12]:
```
cols=['Overview','Original_Language','Poster_Url']
df.drop(cols,axis=1,inplace=True)
df.columns
```

Out[12]:
```
Index(['Release_Date', 'Title', 'Popularity', 'Vote_Count', 'Vote_Average',
       'Genre',
       dtype='object')
```

In [13]:
```
df.head()
```

Out[13]:

	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	3627.658	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	1750	7.0	Action, Adventure, Thriller, War

categorizing vote_avg column

In [14]:
```
def categorize_col(df,col,labels):
    df[col]=pd.to_numeric(df[col],errors='coerce')

    edges=[df[col].min(),
           df[col].quantile(0.25),
           df[col].quantile(0.50),
           df[col].quantile(0.75),
           df[col].max()+0.01]

    df[col]=pd.cut(df[col],bins=edges,labels=labels,duplicates='drop')
    return df
```

In [15]:
```
labels=['non_popular','below_avg','average','popular']

categorize_col(df, 'Vote_Average',labels)

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

Out[15]:
```
['popular', 'below_avg', 'average', 'non_popular', NaN]
Categories (4, object): ['non_popular' < 'below_avg' < 'average' < 'popular']
```

In [16]:
```
df.head()
```

Out[16]:

	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	3627.658	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	1750	average	Action, Adventure, Thriller, War

In [17]:
```
df['Vote_Average'].value_counts()
```

Out[17]:
```
Vote_Average
non_popular    2457
popular        2457
average        2412
below_avg      2398
Name: count, dtype: int64
```

In [18]:
```
df.dropna(inplace=True)

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

Out[18]:
```
Release_Date    0
Title           0
Popularity      0
Vote_Count      0
Vote_Average    0
Genre           0
dtype: int64
```

In [19]:
```
df.head()
```

Out[19]:

	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	3627.658	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	1750	average	Action, Adventure, Thriller, War

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

In [20]:
```
df['Genre'].str.split(', ')

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

Out[20]:

	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	3627.658	1151	popular	Crime
4	2022	The Batman	3627.658	1151	popular	Mystery

In [21]:
```
df.dtypes
```

Out[21]:
```
Release_Date       int32
Title             object
Popularity       float64
Vote_Count         int64
Vote_Average    category
Genre             object
dtype: object
```

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	3627.658	1151	popular	Crime
4	2022	The Batman	3627.658	1151	popular	Mystery

casting col into category

In [23]:
```
df['Genre'].astype('category')

df['Genre'].dtypes
```

In [24]:
```
CategoriesDtype(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 [25]:
```
df.isna().sum()
```

Out[25]:
```
Release_Date     100
Title            9615
Popularity       9615
Vote_Count       9615
Vote_Average        4
Genre             19
dtype: int64
```

data visualization

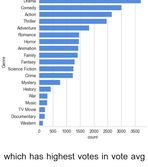
In [26]:
```
sns.set_style('whitegrid')
```

what is the most frequent genre of movies released in netflix?

In [27]:
```
df['Genre'].describe()
```

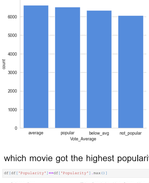
Out[27]:
```
count    25152
unique      19
top      Drama
freq      3715
Name: Genre, dtype: object
```

In [28]:
```
sns.catplot(y='Genre',data=df,kind='count',
            order=df['Genre'].value_counts().index,
            color='#1E7C7C')
plt.title('Genre column distribution')
plt.show()
```



which has highest votes in vote avg column?

In [29]:
```
sns.catplot(x='Vote_Average',data=df,kind='count',
            order=df['Vote_Average'].value_counts().index,
            color='#1E7C7C')
plt.title('Vote distribution')
plt.show()
```



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

In [30]:
```
df.df['Popularity'].loc[df['Popularity'].max()]
```

Out[30]:

	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

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

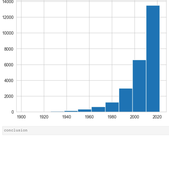
In [31]:
```
df.df['Popularity'].loc[df['Popularity'].min()]
```

Out[31]:

	Release_Date	Title	Popularity	Vote_Count	Vote_Average	Genre
25946	2021	The United States vs. Billie Holiday	13.354	152	average	Music
25947	2021	The United States vs. Billie Holiday	13.354	152	average	Drama
25948	2021	The United States vs. Billie Holiday	13.354	152	average	History
25949	1984	Threads	13.354	186	popular	War
25950	1984	Threads	13.354	186	popular	Drama
25951	1984	Threads	13.354	186	popular	Science Fiction

which year has the most filmmed movies?

In [32]:
```
df['Release_Date'].hist()
plt.title('Release date column distribution')
plt.show()
```



In [33]:
```
conclusion
```