```python
# importing lib.
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns

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

```
     Release_Date                                 Title  \
0      2021-12-15               Spider-Man: No Way Home   
1      2022-03-01                            The Batman   
2      2022-02-25                               No Exit   
3      2021-11-24                               Encanto   
4      2021-12-22                        The King's Man   
...           ...                                   ...   
9822   1973-10-15                              Badlands   
9823   2020-10-01                      Violent Delights   
9824   2016-05-06                          The Offering   
9825   2021-03-31  The United States vs. Billie Holiday   
9826   1984-09-23                               Threads   

                                               Overview  Popularity  \
0     Peter Parker is unmasked and no longer able to...    5083.954   
1     In his second year of fighting crime, Batman u...    3827.658   
2     Stranded at a rest stop in the mountains durin...    2618.087   
3     The tale of an extraordinary family, the Madri...    2402.201   
4     As a collection of history's worst tyrants and...    1895.511   
...                                                 ...         ...   
9822  A dramatization of the Starkweather-Fugate kil...      13.357   
9823  A female vampire falls in love with a man she ...      13.356   
9824  When young and successful reporter Jamie finds...      13.355   
9825  Billie Holiday spent much of her career being ...      13.354   
9826  Documentary style account of a nuclear holocau...      13.354   

      Vote_Count  Vote_Average Original_Language  \
0           8940           8.3                en   
1           1151           8.1                en   
2            122           6.3                en   
3           5076           7.7                en   
4           1793           7.0                en   
...          ...           ...               ...   
9822         896           7.6                en   
9823           8           3.5                es   
9824          94           5.0                en   
9825         152           6.7                en   
9826         186           7.8                en   

                                   Genre  \
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

                                             Poster_Url
0     https://image.tmdb.org/t/p/original/1g0dhYtq4i...
1     https://image.tmdb.org/t/p/original/74xTEgt7R3...
2     https://image.tmdb.org/t/p/original/vDHsLnOWKl...
3     https://image.tmdb.org/t/p/original/4j0PNHkMr5...
4     https://image.tmdb.org/t/p/original/aq4Pwv5Xeu...
...                                                 ...
9822  https://image.tmdb.org/t/p/original/z81rBzHNgi...
9823  https://image.tmdb.org/t/p/original/4b6HY7rud6...
9824  https://image.tmdb.org/t/p/original/h4uMM1wOhz...
9825  https://image.tmdb.org/t/p/original/vEzkxuE2sJ...
9826  https://image.tmdb.org/t/p/original/lBhU4U9Eeh...

[9827 rows x 9 columns]
```

```
df.head()
```

```
  Release_Date                    Title  \
0   2021-12-15  Spider-Man: No Way Home
1   2022-03-01               The Batman
2   2022-02-25                  No Exit
3   2021-11-24                  Encanto
4   2021-12-22           The King's Man

                                            Overview  Popularity
Vote_Count  \
0  Peter Parker is unmasked and no longer able to...    5083.954
8940
1  In his second year of fighting crime, Batman u...    3827.658
1151
2  Stranded at a rest stop in the mountains durin...    2618.087
122
3  The tale of an extraordinary family, the Madri...    2402.201
5076
4  As a collection of history's worst tyrants and...    1895.511
1793
```

```
   Vote_Average Original_Language                               Genre  
\
0           8.3                en  Action, Adventure, Science Fiction  

1           8.1                en            Crime, Mystery, Thriller  

2           6.3                en                            Thriller  

3           7.7                en  Animation, Comedy, Family, Fantasy  

4           7.0                en    Action, Adventure, Thriller, War  


                                          Poster_Url  
0  https://image.tmdb.org/t/p/original/1g0dhYtq4i...  
1  https://image.tmdb.org/t/p/original/74xTEgt7R3...  
2  https://image.tmdb.org/t/p/original/vDHsLnOWKl...  
3  https://image.tmdb.org/t/p/original/4j0PNHkMr5...  
4  https://image.tmdb.org/t/p/original/aq4Pwv5Xeu...  
```

```
# 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
```

```
# exploring genres column
df['Genre'].head()
```

```
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
```

```
# check for duplicated rows
df.duplicated().sum()

0

# exploring summary statistics
df.describe()

        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

# casting column a
df['Release_Date'] = pd.to_datetime(df['Release_Date'])
# confirming changes
print(df['Release_Date'].dtypes)

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

df.info()

df.head()
```

Dropping The Columns Like Dropping Overview, Original_Languege

and Poster-Url

```
# making list of column to be dropped
cols = ['Overview', 'Original_Language', 'Poster_Url']

# dropping columns and confirming changes
df.drop(cols, axis = 1, inplace = True)
df.columns

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

```
df.head()

  Release_Date                    Title  Popularity  Vote_Count
Vote_Average  \
0   2021-12-15  Spider-Man: No Way Home    5083.954        8940
popular
1   2021-12-15  Spider-Man: No Way Home    5083.954        8940
popular
2   2021-12-15  Spider-Man: No Way Home    5083.954        8940
popular
3   2022-03-01               The Batman    3827.658        1151
popular
4   2022-03-01               The Batman    3827.658        1151
popular

             Genre
0           Action
1        Adventure
2  Science Fiction
3            Crime
4          Mystery
```

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.

```
def catigorize_col (df, col, labels):

# setting the edges to cut the column accordingly
    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

# define labels for edges
labels = ['not_popular', 'below_avg', 'average', 'popular']
# categorize column based on labels and edges
catigorize_col(df, 'Vote_Average', labels)
# confirming changes
df['Vote_Average'].unique()
```

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

```
df.head()
```

```
  Release_Date                    Title  \
0   2021-12-15  Spider-Man: No Way Home
1   2022-03-01               The Batman
2   2022-02-25                  No Exit
3   2021-11-24                  Encanto
4   2021-12-22           The King's Man

                                            Overview  Popularity
Vote_Count  \
0  Peter Parker is unmasked and no longer able to...    5083.954
8940
1  In his second year of fighting crime, Batman u...    3827.658
1151
2  Stranded at a rest stop in the mountains durin...    2618.087
122
3  The tale of an extraordinary family, the Madri...    2402.201
5076
4  As a collection of history's worst tyrants and...    1895.511
1793

   Vote_Average Original_Language                               Genre
\
0           8.3                en  Action, Adventure, Science Fiction

1           8.1                en            Crime, Mystery, Thriller

2           6.3                en                            Thriller

3           7.7                en  Animation, Comedy, Family, Fantasy

4           7.0                en    Action, Adventure, Thriller, War


                                          Poster_Url
0  https://image.tmdb.org/t/p/original/1g0dhYtq4i...
1  https://image.tmdb.org/t/p/original/74xTEgt7R3...
2  https://image.tmdb.org/t/p/original/vDHsLnOWKl...
3  https://image.tmdb.org/t/p/original/4j0PNHkMr5...
4  https://image.tmdb.org/t/p/original/aq4Pwv5Xeu...
```

```
# exploring column
df['Vote_Average'].value_counts()
```

```
Vote_Average
not_popular    2467
```

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

# dropping NaNs
df.dropna(inplace = True)
# confirming
df.isna().sum()

Release_Date         0
Title                0
Overview             0
Popularity           0
Vote_Count           0
Vote_Average         0
Original_Language    0
Genre                0
Poster_Url           0
dtype: int64

df.head()

  Release_Date                    Title  \
0   2021-12-15  Spider-Man: No Way Home
1   2022-03-01               The Batman
2   2022-02-25                  No Exit
3   2021-11-24                  Encanto
4   2021-12-22           The King's Man

                                            Overview  Popularity
Vote_Count  \
0  Peter Parker is unmasked and no longer able to...    5083.954
8940
1  In his second year of fighting crime, Batman u...    3827.658
1151
2  Stranded at a rest stop in the mountains durin...    2618.087
122
3  The tale of an extraordinary family, the Madri...    2402.201
5076
4  As a collection of history's worst tyrants and...    1895.511
1793

  Vote_Average Original_Language
Genre  \
0      popular                en  Action, Adventure, Science Fiction

1      popular                en            Crime, Mystery, Thriller

2    below_avg                en                            Thriller
```

```
3      popular                en  Animation, Comedy, Family, Fantasy

4      average                en    Action, Adventure, Thriller, War


                                          Poster_Url
0  https://image.tmdb.org/t/p/original/1g0dhYtq4i...
1  https://image.tmdb.org/t/p/original/74xTEgt7R3...
2  https://image.tmdb.org/t/p/original/vDHsLnOWKl...
3  https://image.tmdb.org/t/p/original/4j0PNHkMr5...
4  https://image.tmdb.org/t/p/original/aq4Pwv5Xeu...
```

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

```
# split the strings into lists
df['Genre'] = df['Genre'].str.split(', ')
# explode the lists
df = df.explode('Genre').reset_index(drop=True)
df.head()

  Release_Date                    Title  \
0   2021-12-15  Spider-Man: No Way Home
1   2021-12-15  Spider-Man: No Way Home
2   2021-12-15  Spider-Man: No Way Home
3   2022-03-01               The Batman
4   2022-03-01               The Batman

                                            Overview  Popularity
Vote_Count  \
0  Peter Parker is unmasked and no longer able to...    5083.954
8940
1  Peter Parker is unmasked and no longer able to...    5083.954
8940
2  Peter Parker is unmasked and no longer able to...    5083.954
8940
3  In his second year of fighting crime, Batman u...    3827.658
1151
4  In his second year of fighting crime, Batman u...    3827.658
1151

  Vote_Average Original_Language            Genre  \
0      popular                en           Action
1      popular                en        Adventure
2      popular                en  Science Fiction
3      popular                en            Crime
4      popular                en          Mystery
```

```
                                          Poster_Url
0  https://image.tmdb.org/t/p/original/1g0dhYtq4i...
1  https://image.tmdb.org/t/p/original/1g0dhYtq4i...
2  https://image.tmdb.org/t/p/original/1g0dhYtq4i...
3  https://image.tmdb.org/t/p/original/74xTEgt7R3...
4  https://image.tmdb.org/t/p/original/74xTEgt7R3...
```

```
# casting column into category
df['Genre'] = df['Genre'].astype('category')
# confirming changes
df['Genre'].dtypes
```

```
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)
```

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 25552 entries, 0 to 25551
Data columns (total 9 columns):
 #   Column             Non-Null Count  Dtype
---  ------             --------------  -----
 0   Release_Date       25552 non-null  object
 1   Title              25552 non-null  object
 2   Overview           25552 non-null  object
 3   Popularity         25552 non-null  float64
 4   Vote_Count         25552 non-null  int64
 5   Vote_Average       25552 non-null  category
 6   Original_Language  25552 non-null  object
 7   Genre              25552 non-null  category
 8   Poster_Url         25552 non-null  object
dtypes: category(2), float64(1), int64(1), object(5)
memory usage: 1.4+ MB
```

```
df.nunique()
```

```
Release_Date         5846
Title                9415
Overview             9722
Popularity           8088
Vote_Count           3265
Vote_Average            4
Original_Language      42
Genre                  19
```

```
Poster_Url          9727
dtype: int64
```

Data Visualization

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

```
# setting up seaborn configurations
sns.set_style('whitegrid')
```

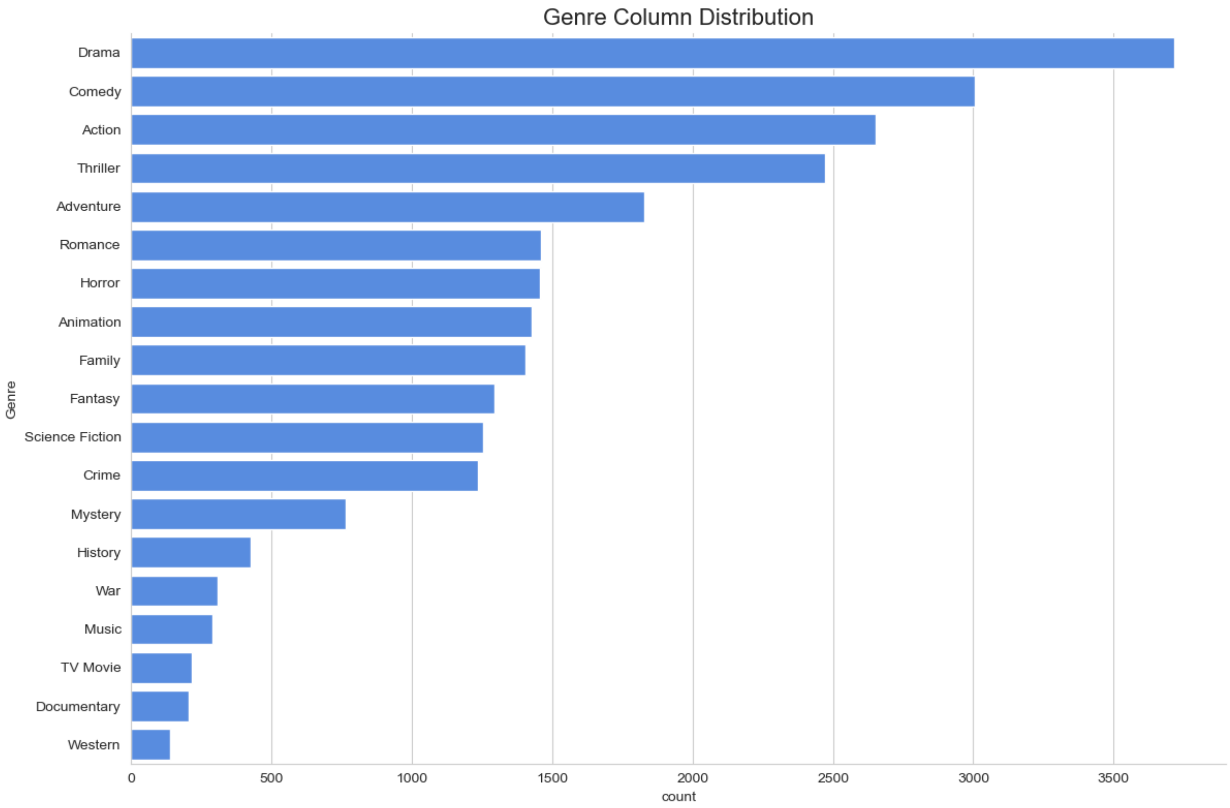
# Q1: What is the most frequent genre of movies released on Netflix ? the dataset?

```
# showing stats. on genre column
df['Genre'].describe()
```

```
count     25552
unique       19
top       Drama
freq       3715
Name: Genre, dtype: object
```

```
# Plotting genre distribution
sns.catplot(
    y='Genre',
    data=df,
    kind='count',
    order=df['Genre'].value_counts().index,
    color='#4287f5',
    height=8,        # height of the plot
    aspect=1.5       # aspect ratio (width = height * aspect)
)

plt.title('Genre Column Distribution', fontsize=16)
plt.tight_layout()
plt.show()
```



Q2 Which Has Highest votes in vote avg column?

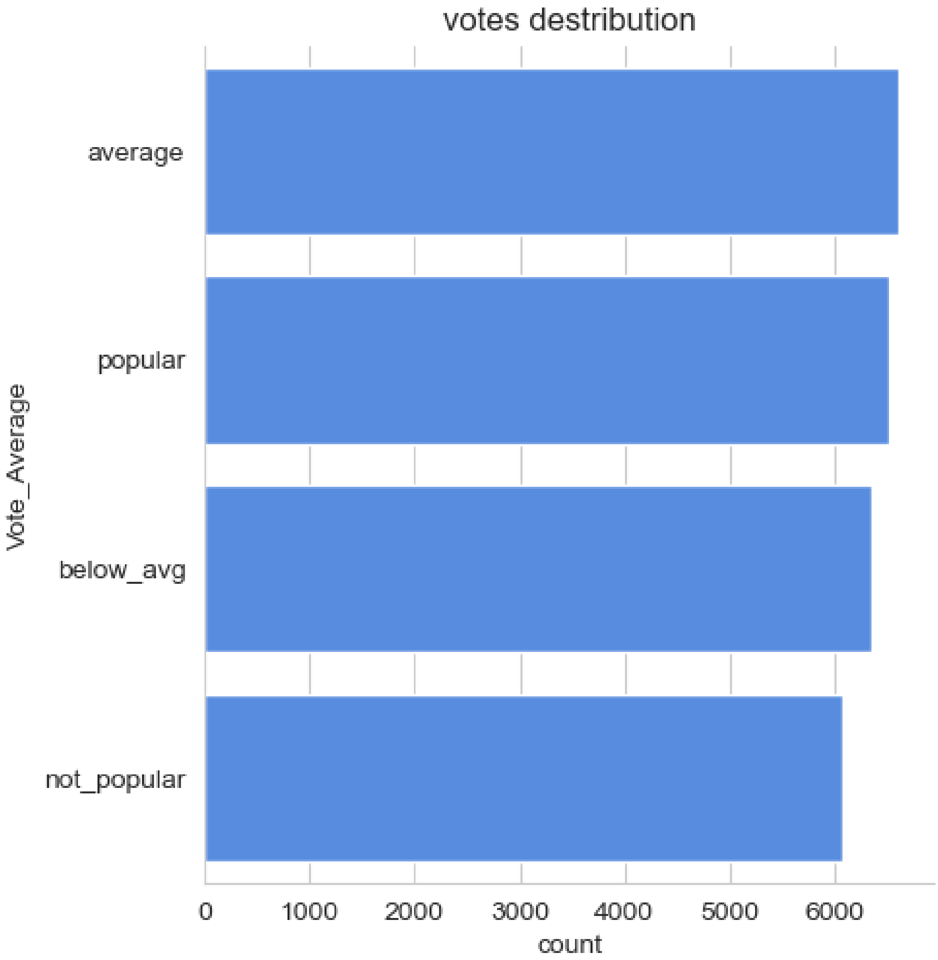
```
df.head()

  Release_Date                    Title  Popularity  Vote_Count
Vote_Average  \
0   2021-12-15  Spider-Man: No Way Home    5083.954        8940
popular
1   2021-12-15  Spider-Man: No Way Home    5083.954        8940
popular
2   2021-12-15  Spider-Man: No Way Home    5083.954        8940
popular
3   2022-03-01               The Batman    3827.658        1151
popular
4   2022-03-01               The Batman    3827.658        1151
popular

             Genre
0           Action
```

```
1         Adventure
2   Science Fiction
3             Crime
4           Mystery

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



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

```
df.head(2)

  Release_Date                   Title  Popularity  Vote_Count
Vote_Average  \
0   2021-12-15  Spider-Man: No Way Home    5083.954        8940
popular
1   2021-12-15  Spider-Man: No Way Home    5083.954        8940
popular

       Genre
0     Action
1  Adventure

# checking max popularity in dataset
df[df['Popularity'] == df['Popularity'].max()]

  Release_Date                   Title  Popularity  Vote_Count
Vote_Average  \
0   2021-12-15  Spider-Man: No Way Home    5083.954        8940
popular
1   2021-12-15  Spider-Man: No Way Home    5083.954        8940
popular
2   2021-12-15  Spider-Man: No Way Home    5083.954        8940
popular

             Genre
0           Action
1        Adventure
2  Science Fiction
```

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

```
# checking max popularity in dataset
df[df['Popularity'] == df['Popularity'].min()]
```

```
     Release_Date                                              Title  \
9825   2021-03-31  The United States vs. Billie Holiday
9826   1984-09-23                                            Threads

                                               Overview  Popularity  \
9825  Billie Holiday spent much of her career being ...      13.354
9826  Documentary style account of a nuclear holocau...      13.354

      Vote_Count  Vote_Average Original_Language
Genre  \
9825         152           6.7                en        Music, Drama,
History
9826         186           7.8                en  War, Drama, Science
Fiction

                                             Poster_Url
9825  https://image.tmdb.org/t/p/original/vEzkxuE2sJ...
9826  https://image.tmdb.org/t/p/original/lBhU4U9Eeh...
```

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`.
```

THANK YOU