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

```
df=pd.read_csv('mymoviedb (1).csv',lineterminator='\n')
```

```
df.head()
```

```
  Release_Date                    Title  Popularity  Vote_Count  Vote_Average  \
0   2021-12-15  Spider-Man: No Way Home    5083.954        8940           8.3
1   2022-03-01               The Batman    3827.658        1151           8.1
2   2022-02-25                  No Exit    2618.087         122           6.3
3   2021-11-24                  Encanto    2402.201        5076           7.7
4   2021-12-22           The King's Man    1895.511        1793           7.0

                              Genre
0  Action, Adventure, Science Fiction
1            Crime, Mystery, Thriller
2                            Thriller
3  Animation, Comedy, Family, Fantasy
4    Action, Adventure, Thriller, War
```

```
df.duplicated().sum()
```

```
np.int64(0)
```

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 9827 entries, 0 to 9826
Data columns (total 6 columns):
 #   Column        Non-Null Count  Dtype
---  ------        --------------  -----
 0   Release_Date  9827 non-null   object
 1   Title         9827 non-null   object
 2   Popularity    9827 non-null   float64
 3   Vote_Count    9827 non-null   int64
 4   Vote_Average  9827 non-null   float64
 5   Genre         9827 non-null   object
dtypes: float64(2), int64(1), object(3)
memory usage: 460.8+ KB
```

```
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
```

• Exploration Summary • we have a dataframe consisting of 9827 rows and 9 columns. • our dataset looks a bit tidy with no NaNs nor duplicated values. • Release_Date column needs to be casted into date time and to extract only the • Overview, Original_Languege and Poster-Url wouldn't be so useful during analys • there is noticable outliers in Popularity column • Vote_Average bettter be categorised for proper analysis. • Genre column has comma saperated values and white spaces that needs to be hand

```
cols=['Overview','Original_Language','Poster_Url']
```

```
df.drop(cols,axis=1,inplace=True)
```

```
df.head()
```

```
  Release_Date                    Title  Popularity  Vote_Count  Vote_Average  \
0   2021-12-15  Spider-Man: No Way Home    5083.954        8940           8.3
1   2022-03-01               The Batman    3827.658        1151           8.1
2   2022-02-25                  No Exit    2618.087         122           6.3
3   2021-11-24                  Encanto    2402.201        5076           7.7
4   2021-12-22           The King's Man    1895.511        1793           7.0

                              Genre
0  Action, Adventure, Science Fiction
1            Crime, Mystery, Thriller
2                            Thriller
3  Animation, Comedy, Family, Fantasy
4    Action, Adventure, Thriller, War
```

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

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

```
df.head()
```

```
   Release_Date                    Title  Popularity  Vote_Count  \
0          2021  Spider-Man: No Way Home    5083.954        8940
1          2022               The Batman    3827.658        1151
2          2022                  No Exit    2618.087         122
3          2021                  Encanto    2402.201        5076
4          2021           The King's Man    1895.511        1793

   Vote_Average                              Genre
0           8.3  Action, Adventure, Science Fiction
1           8.1            Crime, Mystery, Thriller
2           6.3                            Thriller
```

```
3           7.7  Animation, Comedy, Family, Fantasy
4           7.0    Action, Adventure, Thriller, War
```

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 catogarize_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
```

```
labels=['not_popular','below_average','average','popular']
```

```
catogarize_col(df,'Vote_Average',labels)
df['Vote_Average'].unique()
```

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

```
df.head()
```

```
   Release_Date                    Title  Popularity  Vote_Count  \
0          2021  Spider-Man: No Way Home    5083.954        8940
1          2022               The Batman    3827.658        1151
2          2022                  No Exit    2618.087         122
3          2021                  Encanto    2402.201        5076
4          2021           The King's Man    1895.511        1793

    Vote_Average                                Genre
0        popular  Action, Adventure, Science Fiction
1        popular            Crime, Mystery, Thriller
2  below_average                             Thriller
3        popular  Animation, Comedy, Family, Fantasy
4        average    Action, Adventure, Thriller, War
```

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

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

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

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

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

```
df.head()
```

```
   Release_Date                    Title  Popularity  Vote_Count  \
0          2021  Spider-Man: No Way Home    5083.954        8940
1          2022               The Batman    3827.658        1151
2          2022                  No Exit    2618.087         122
3          2021                  Encanto    2402.201        5076
4          2021            The King's Man    1895.511        1793

    Vote_Average                                Genre
0        popular  Action, Adventure, Science Fiction
1        popular            Crime, Mystery, Thriller
2  below_average                             Thriller
3        popular  Animation, Comedy, Family, Fantasy
4        average    Action, Adventure, Thriller, War
```

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

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

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

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

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

```
CategoricalDtype(categories=[' Action', ' Adventure', ' Animation', ' Comedy', ' Crime',
                  ' Documentary', ' Drama', ' Family', ' Fantasy', ' History',
                  ' Horror', ' Music', ' Mystery', ' Romance',
                  ' Science Fiction', ' TV Movie', ' Thriller', ' War',
                  ' Western', '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 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: 750.3+ KB
```

```
df.nunique()
```

```
Release_Date     100
Title           9415
Popularity      8088
Vote_Count      3265
Vote_Average       4
Genre             38
dtype: int64
```

```
df.head()
```

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

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

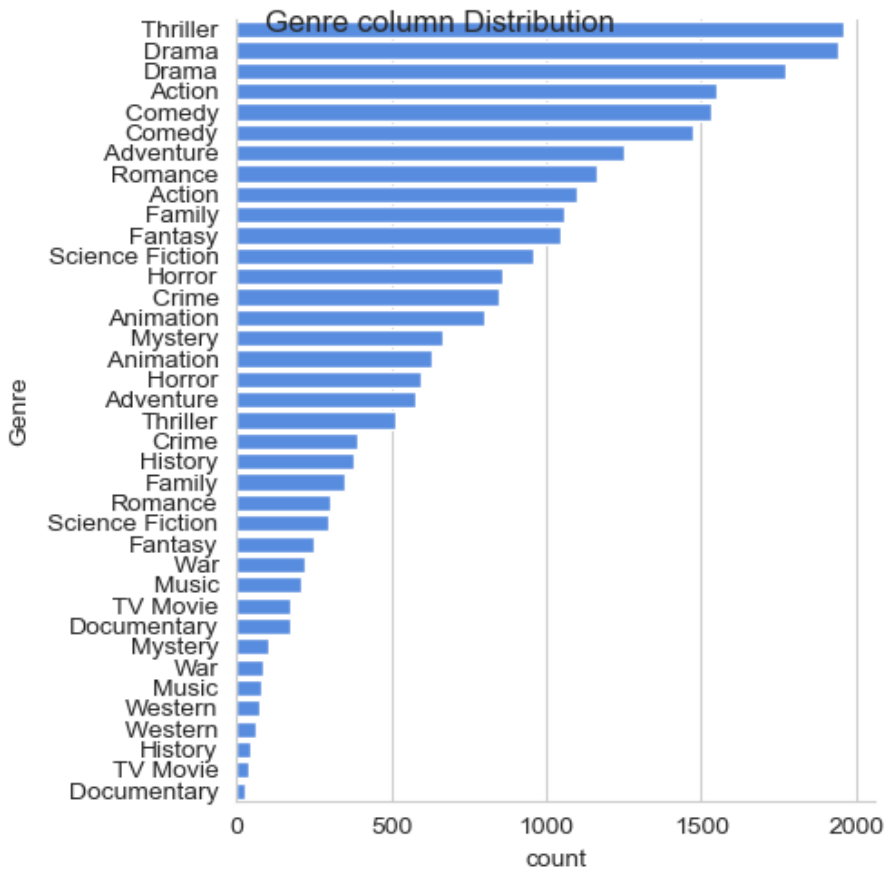
Data visualization

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

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

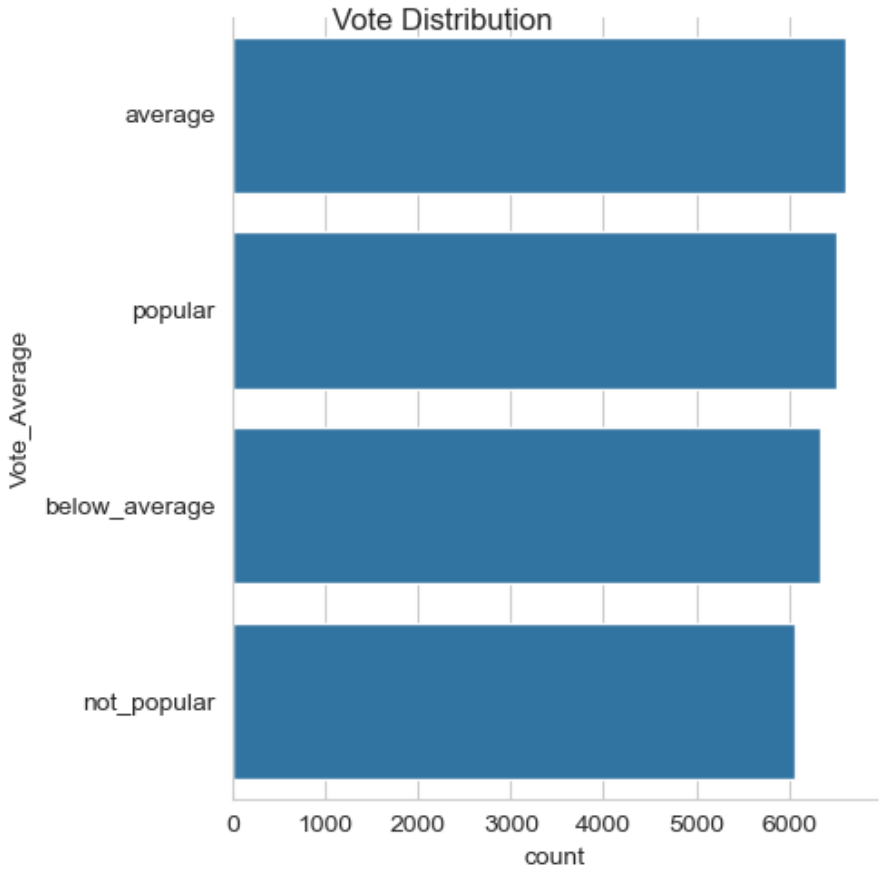
#what is the most frequent genre in movie released on netflix?

```
g=(sns.catplot(y='Genre',data=df,kind='count',
            order= df['Genre'].value_counts().index,
            color='#4287f5'))
g.fig.suptitle('Genre column Distribution')
plt.show()
```




#which has highest vote in vote_Average column?

```
g=(sns.catplot(y='Vote_Average',data=df,kind='count',
               order=df['Vote_Average'].value_counts().index))
g.fig.suptitle('Vote Distribution')
plt.show()
```




#which movie got highest popularity and which genre it is?

```
df.head()
```

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

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

```
df[df['Popularity']==df['Popularity'].max()]
```

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

              Genre
0            Action
1         Adventure
2   Science Fiction
```

#which movie got lowest popularity and which genre it is?

```
df[df['Popularity']==df['Popularity'].min()]
```

```
       Release_Date                                Title  Popularity  \
25546          2021  The United States vs. Billie Holiday      13.354
25547          2021  The United States vs. Billie Holiday      13.354
25548          2021  The United States vs. Billie Holiday      13.354
25549          1984                               Threads      13.354
25550          1984                               Threads      13.354
25551          1984                               Threads      13.354

       Vote_Count Vote_Average             Genre
25546         152      average             Music
25547         152      average             Drama
25548         152      average           History
25549         186      popular               War
25550         186      popular             Drama
25551         186      popular   Science Fiction
```

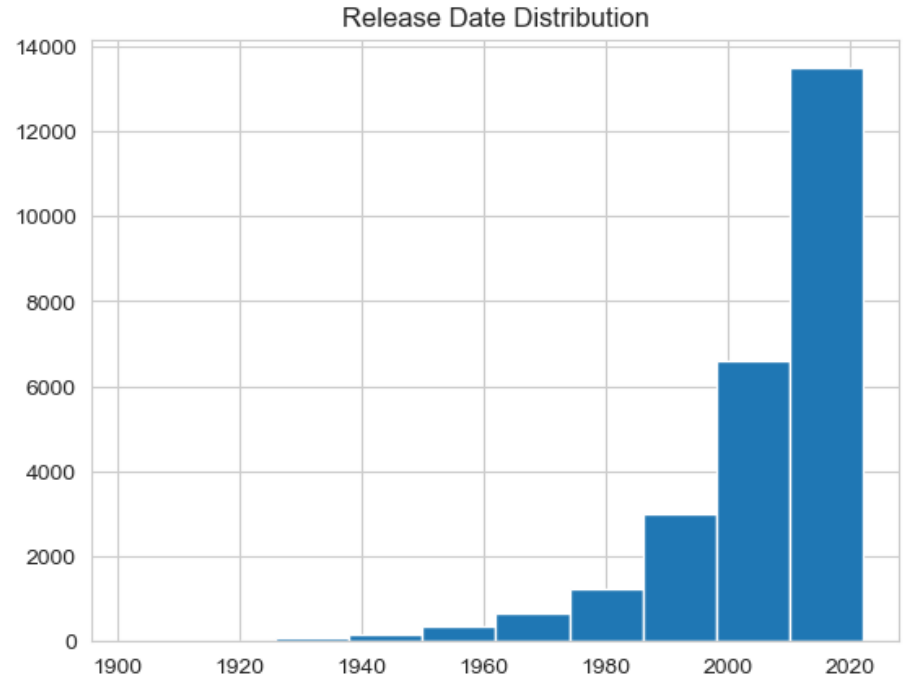
#which year has most filmmed movies?

```
df.head()
```

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

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

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

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