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

df = pd.read_csv("file:///D:/all/mymoviedb.csv", lineterminator='\n')
df.head()
```

```
  Release_Date                    Title  \
0   12/15/2021  Spider-Man: No Way Home
1     3/1/2022               The Batman
2    2/25/2022                  No Exit
3   11/24/2021                  Encanto
4   12/22/2021           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                en    Action, Adventure, Thriller, War


                                        Poster_Url\r
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...
```

```python
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 9837 entries, 0 to 9836
Data columns (total 9 columns):
 #   Column             Non-Null Count  Dtype  
---  ------             --------------  -----  
 0   Release_Date       9837 non-null   object 
 1   Title              9828 non-null   object 
 2   Overview           9828 non-null   object 
 3   Popularity         9827 non-null   float64
 4   Vote_Count         9827 non-null   object 
 5   Vote_Average       9827 non-null   object 
 6   Original_Language  9827 non-null   object 
 7   Genre              9826 non-null   object 
        9837 non-null   object 
dtypes: float64(1), object(8)
memory usage: 691.8+ 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
count  9827.000000
mean     40.320570
std     108.874308
min       7.100000
25%      16.127500
50%      21.191000
75%      35.174500
max    5083.954000
```

```
# Data Cleaning
#Casting Release_Date column and extracing year values

df.head()
```

```
  Release_Date                    Title  \
0   12/15/2021  Spider-Man: No Way Home
1     3/1/2022               The Batman
2    2/25/2022                  No Exit
3   11/24/2021                  Encanto
4   12/22/2021           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                en    Action, Adventure, Thriller, War


                                          Poster_Url\r
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...
```

```
# casting column a
df["Release_Date"] = pd.to_datetime(df["Release_Date"],
format="%m/%d/%Y", errors='coerce')
# confirming changes
print(df['Release_Date'].dtypes)
```

```
datetime64[ns]
```

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

```
dtype('float64')
```

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 9837 entries, 0 to 9836
Data columns (total 9 columns):
 #   Column             Non-Null Count  Dtype  
---  ------             --------------  -----  
 0   Release_Date       9827 non-null   float64
 1   Title              9828 non-null   object 
 2   Overview           9828 non-null   object 
 3   Popularity         9827 non-null   float64
 4   Vote_Count         9827 non-null   object 
 5   Vote_Average       9827 non-null   object 
 6   Original_Language  9827 non-null   object 
 7   Genre              9826 non-null   object 
        9837 non-null   object 
dtypes: float64(2), object(7)
memory usage: 691.8+ KB
```

```
df.head()
```

```
   Release_Date                    Title  \
0        2021.0  Spider-Man: No Way Home   
1        2022.0               The Batman   
2        2022.0                  No Exit   
3        2021.0                  Encanto   
4        2021.0           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                en    Action, Adventure, Thriller, War


                                         Poster_Url\r
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...
```

```
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

df.head()
```

```
   Release_Date                    Title  Popularity Vote_Count
Vote_Average  \
0        2021.0  Spider-Man: No Way Home    5083.954       8940
8.3
1        2022.0               The Batman    3827.658       1151
8.1
2        2022.0                  No Exit    2618.087        122
6.3
3        2021.0                  Encanto    2402.201       5076
7.7
4        2021.0           The King's Man    1895.511       1793
7

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

                                         Poster_Url\r
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
6.4          435
6.3          429
6.5          427
6.8          423
6.7          420
            ...
9.2            1
1.5            1
3.1            1
Animation      1
10             1
Name: count, Length: 75, dtype: int64
```

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

```
Release_Date    0
Title           0
Popularity      0
Vote_Count      0
Vote_Average    0
Genre           0
Poster_Url\r    0
dtype: int64
```

```
df.head()
```

```
   Release_Date                    Title  Popularity Vote_Count
Vote_Average  \
0        2021.0  Spider-Man: No Way Home    5083.954       8940
8.3
1        2022.0               The Batman    3827.658       1151
8.1
2        2022.0                  No Exit    2618.087        122
6.3
3        2021.0                  Encanto    2402.201       5076
7.7
4        2021.0           The King's Man    1895.511       1793
7

                                Genre  \
0  Action, Adventure, Science Fiction
```

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

                                         Poster_Url\r
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...
```

split genres into a list and thenexplode our dataframe to have only one genre per row for each 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  Popularity Vote_Count
Vote_Average  \
0        2021.0  Spider-Man: No Way Home    5083.954       8940
8.3
1        2021.0  Spider-Man: No Way Home    5083.954       8940
8.3
2        2021.0  Spider-Man: No Way Home    5083.954       8940
8.3
3        2022.0               The Batman    3827.658       1151
8.1
4        2022.0               The Batman    3827.658       1151
8.1

             Genre                                       Poster_Url\r

0           Action  https://image.tmdb.org/t/p/original/1g0dhYtq4i...

1        Adventure  https://image.tmdb.org/t/p/original/1g0dhYtq4i...

2  Science Fiction  https://image.tmdb.org/t/p/original/1g0dhYtq4i...

3            Crime  https://image.tmdb.org/t/p/original/74xTEgt7R3...

4          Mystery  https://image.tmdb.org/t/p/original/74xTEgt7R3...


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

```python
# 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)
```

```python
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 25792 entries, 0 to 25791
Data columns (total 7 columns):
 #   Column        Non-Null Count  Dtype
---  ------        --------------  -----
 0   Release_Date  25792 non-null  float64
 1   Title         25792 non-null  object
 2   Popularity    25792 non-null  float64
 3   Vote_Count    25792 non-null  object
 4   Vote_Average  25792 non-null  object
 5   Genre         25792 non-null  category
   25792 non-null  object
dtypes: category(1), float64(2), object(4)
memory usage: 1.2+ MB
```

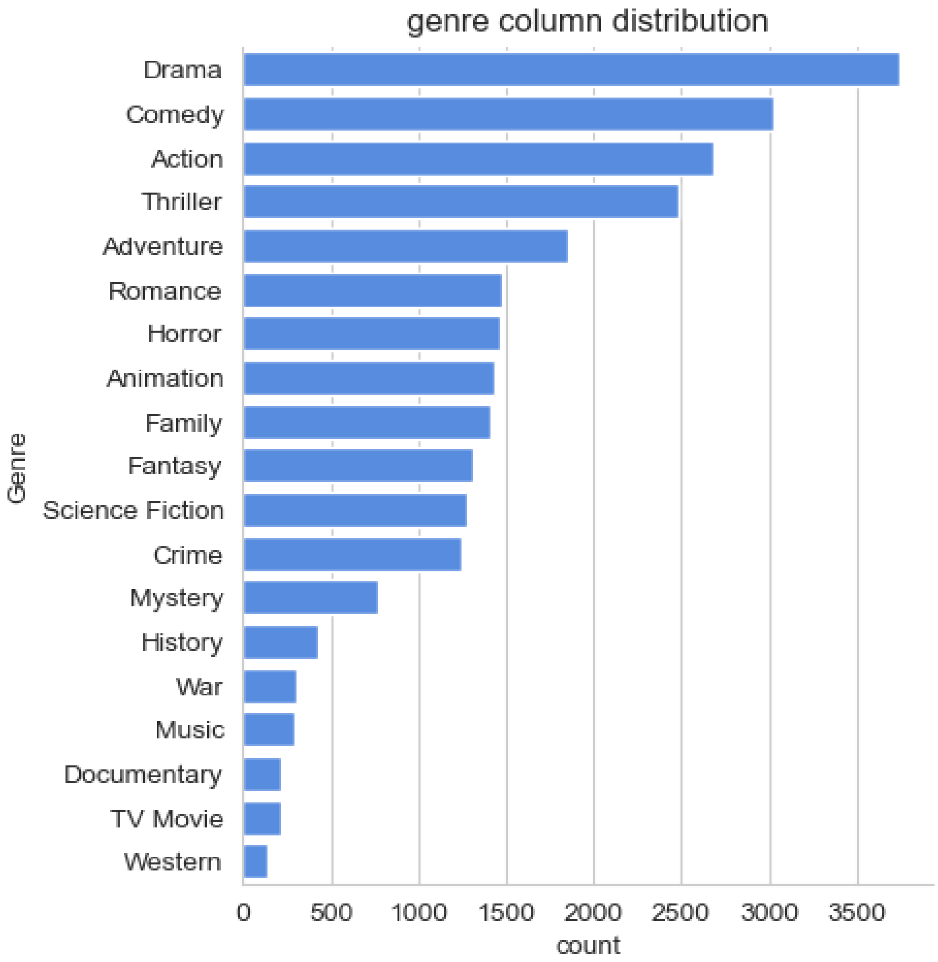
```python
df.nunique()
```

```
Release_Date     102
Title           9512
Popularity      8159
Vote_Count      3266
Vote_Average      74
Genre             19
Poster_Url\r    9826
dtype: int64
```

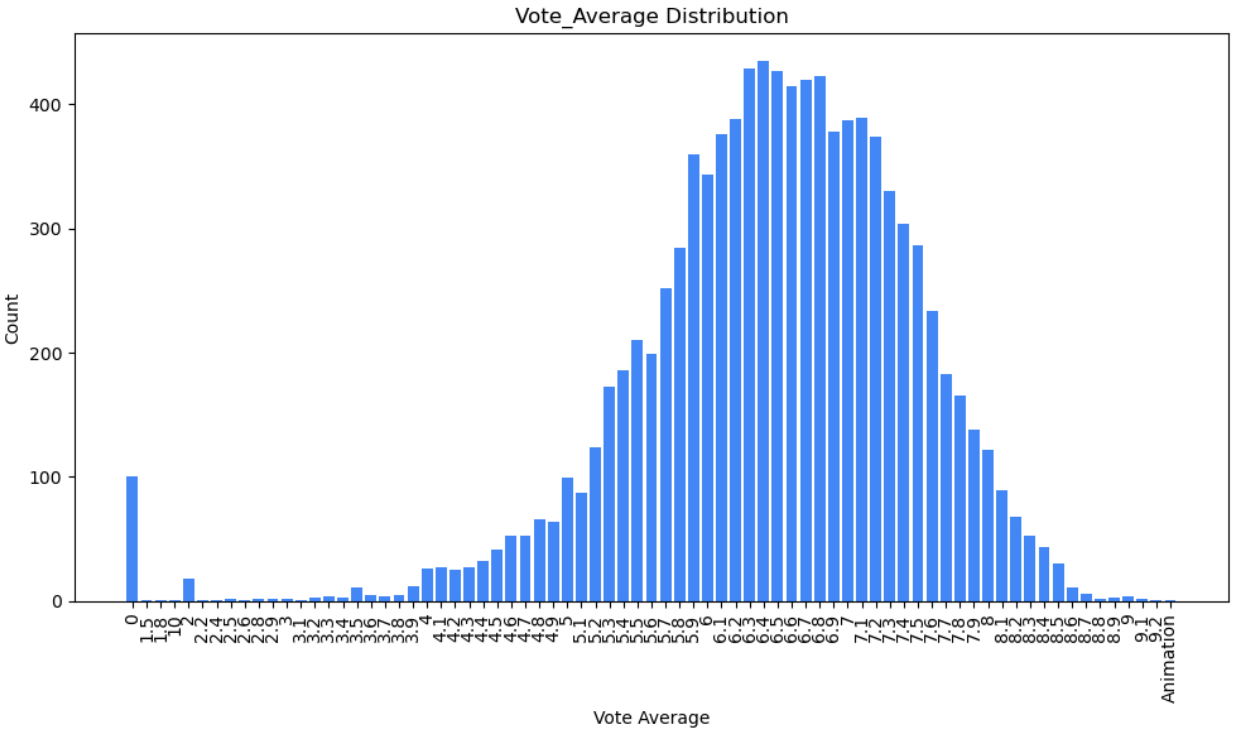
```python
# setting up seaborn configurations
sns.set_style('whitegrid')

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

```
count     25792
unique       19
top       Drama
freq       3744
Name: Genre, dtype: object
```

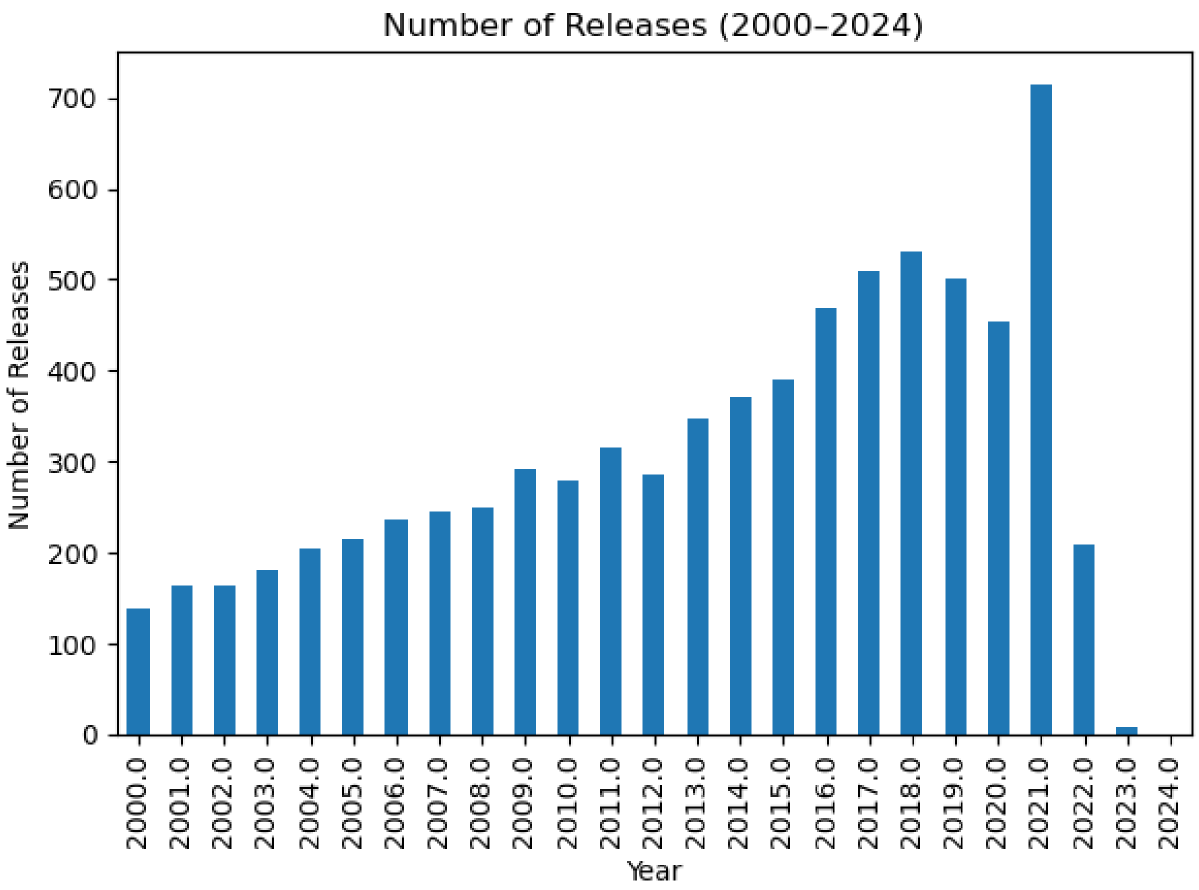
```python
# visualizing genre column
sns.catplot(y = 'Genre', data = df, kind = 'count',
 order = df['Genre'].value_counts().index,
 color = '#4287f5')
plt.title('genre column distribution')
plt.show()
```



```python
import matplotlib.pyplot as plt

# Count occurrences of each Vote_Average
vote_counts = df['Vote_Average'].value_counts().sort_index()

# Plot bar chart
plt.figure(figsize=(10, 6))
plt.bar(vote_counts.index.astype(str), vote_counts.values,
color='#4287f5')
plt.title('Vote_Average Distribution')
plt.xlabel('Vote Average')
plt.ylabel('Count')
plt.xticks(rotation=90)
plt.tight_layout()
plt.show()
```



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

  Release_Date                    Title  \
0   12/15/2021  Spider-Man: No Way Home

                                            Overview  Popularity
Vote_Count  \
0  Peter Parker is unmasked and no longer able to...    5083.954
8940

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


                                        Poster_Url\r
0  https://image.tmdb.org/t/p/original/1g0dhYtq4i...

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

         Release_Date   Title Overview  Popularity Vote_Count
Vote_Average  \
1115   - Magic Tricks  61.328       35         7.1         en
Animation

                                      Original_Language Genre
Poster_Url\r
1115  https://image.tmdb.org/t/p/original/6iXYe7AkQ1...   NaN
\r

# Ensure datetime format
df['Release_Date'] = pd.to_datetime(df['Release_Date'],
errors='coerce')

# Extract year
df['Release_Year'] = df['Release_Date'].dt.year

# Filter for year range
filtered = df[(df['Release_Year'] >= 2000) & (df['Release_Year'] <=
2024)]

# Count releases per year
release_counts = filtered['Release_Year'].value_counts().sort_index()

# Plot
release_counts.plot(kind='bar')
plt.title('Number of Releases (2000–2024)')
plt.xlabel('Year')
plt.ylabel('Number of Releases')
plt.xticks(rotation=90)
plt.tight_layout()
plt.show()
```



Number of Releases (2000–2024)
Number of Releases
0
100
200
300
400
500
600
700
2000.0
2001.0
2002.0
2003.0
2004.0
2005.0
2006.0
2007.0
2008.0
2009.0
2010.0
2011.0
2012.0
2013.0
2014.0
2015.0
2016.0
2017.0
2018.0
2019.0
2020.0
2021.0
2022.0
2023.0
2024.0
Year

Conclusion