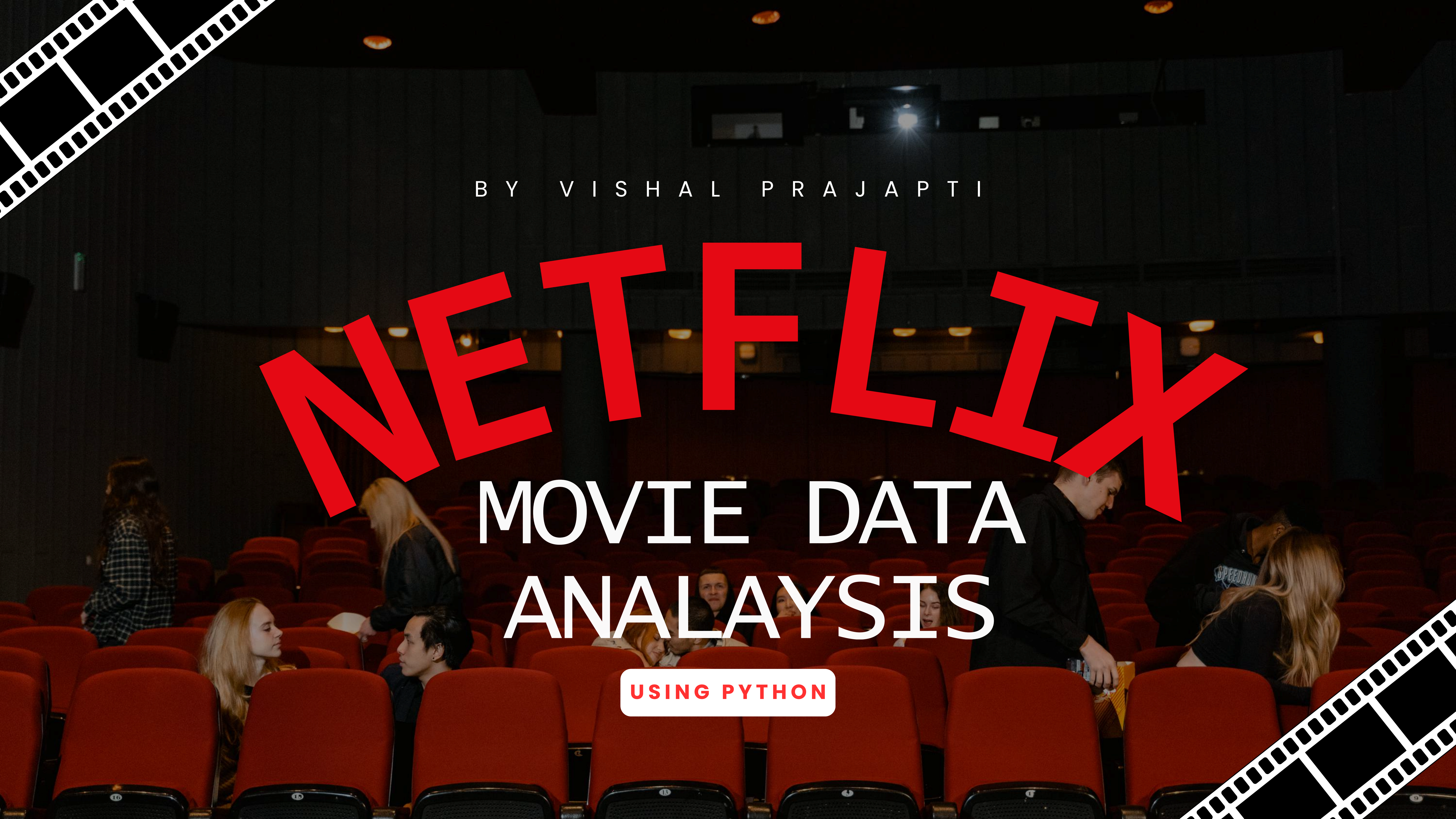


BY VISHAL PRAJAPTI

NETFLIX

MOVIE DATA ANALAYSIS

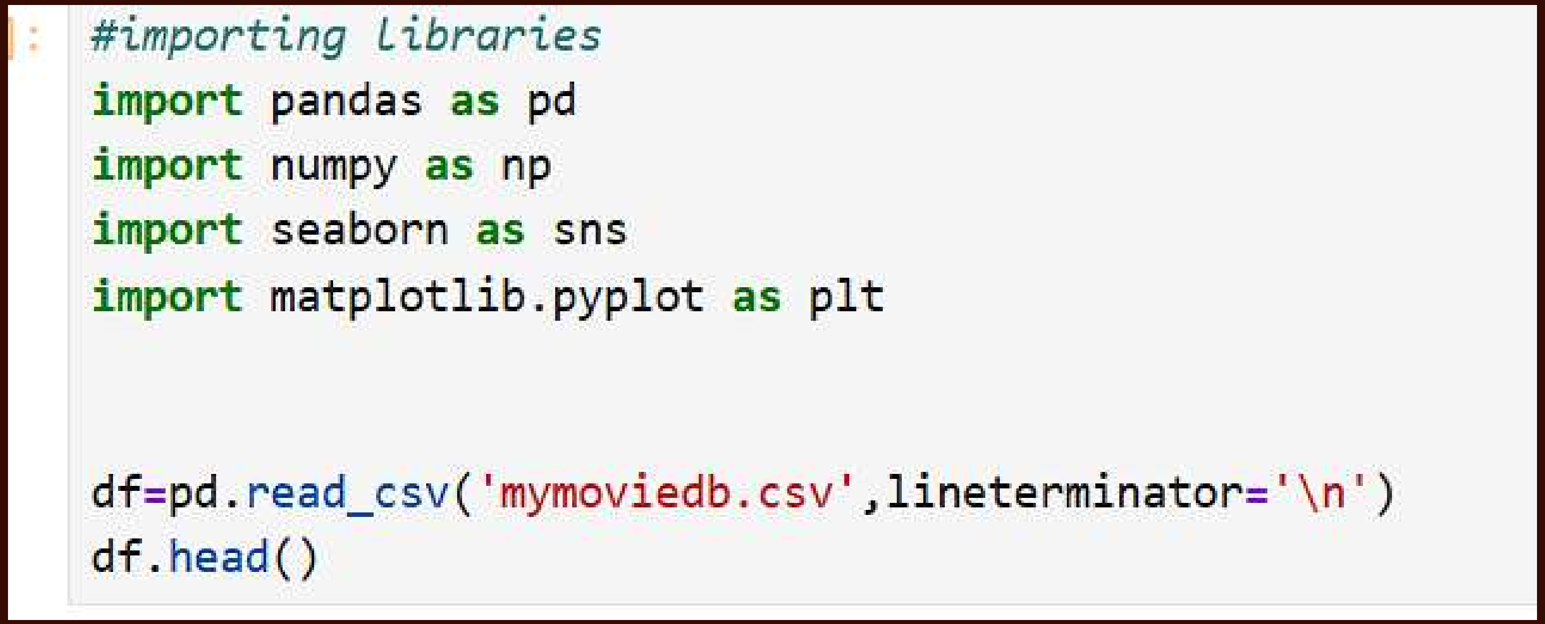
USING PYTHON

Netflix Movie Data Analysis Project

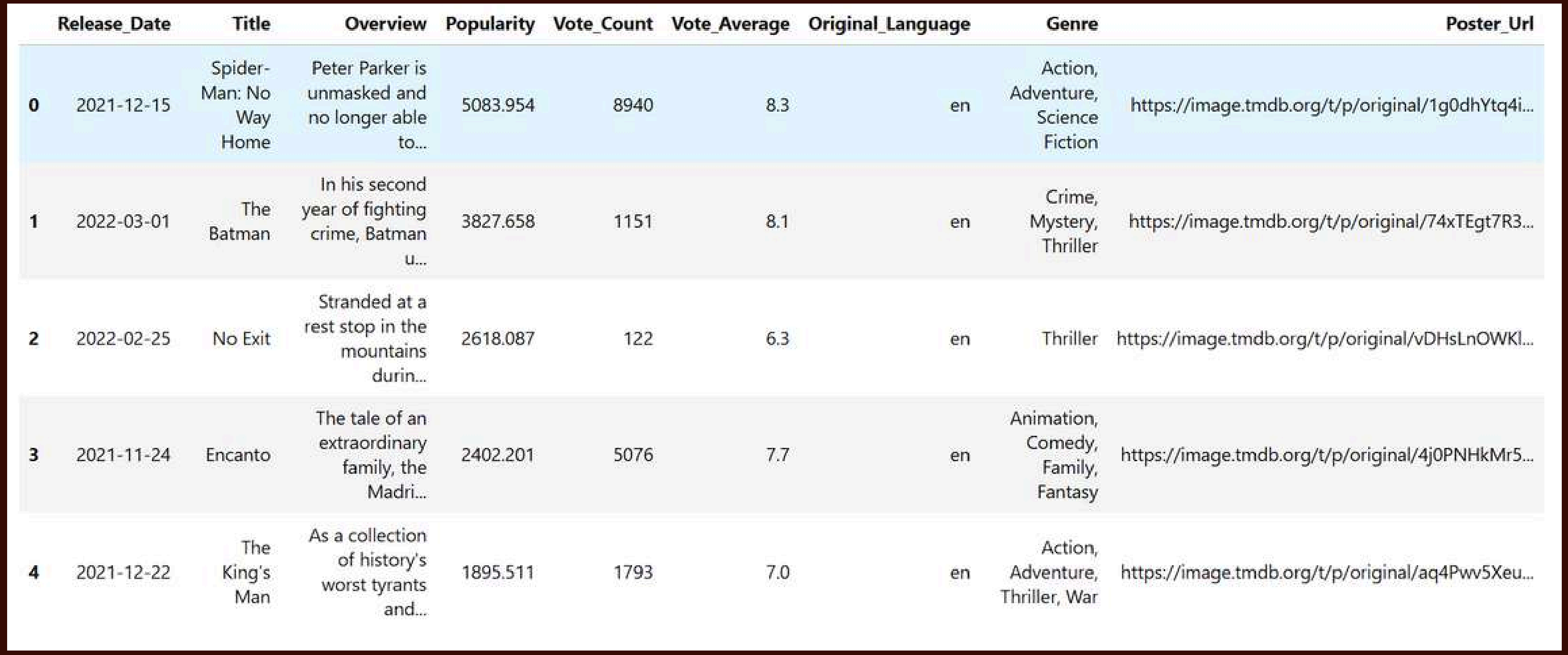
Netflix is known for its work in data science, AI, and ML, particularly for building strong recommendation models and algorithms that understand customer behavior and patterns. Suppose you are working in a data-driven job role, and you have a dataset of more than 9,000 movies. You need to solve the following questions to help the company make informed business decisions accordingly.
ph text

1) What is the most frequent genre of movies released on Netflix?
2) Which has highest votes in vote avg column?
3) What movie got the highest popularity? what's its genre?
4) What movie got the lowest popularity? what's its genre?
5) Which year has the most filmmed movies?

DATA PREPROCESSING

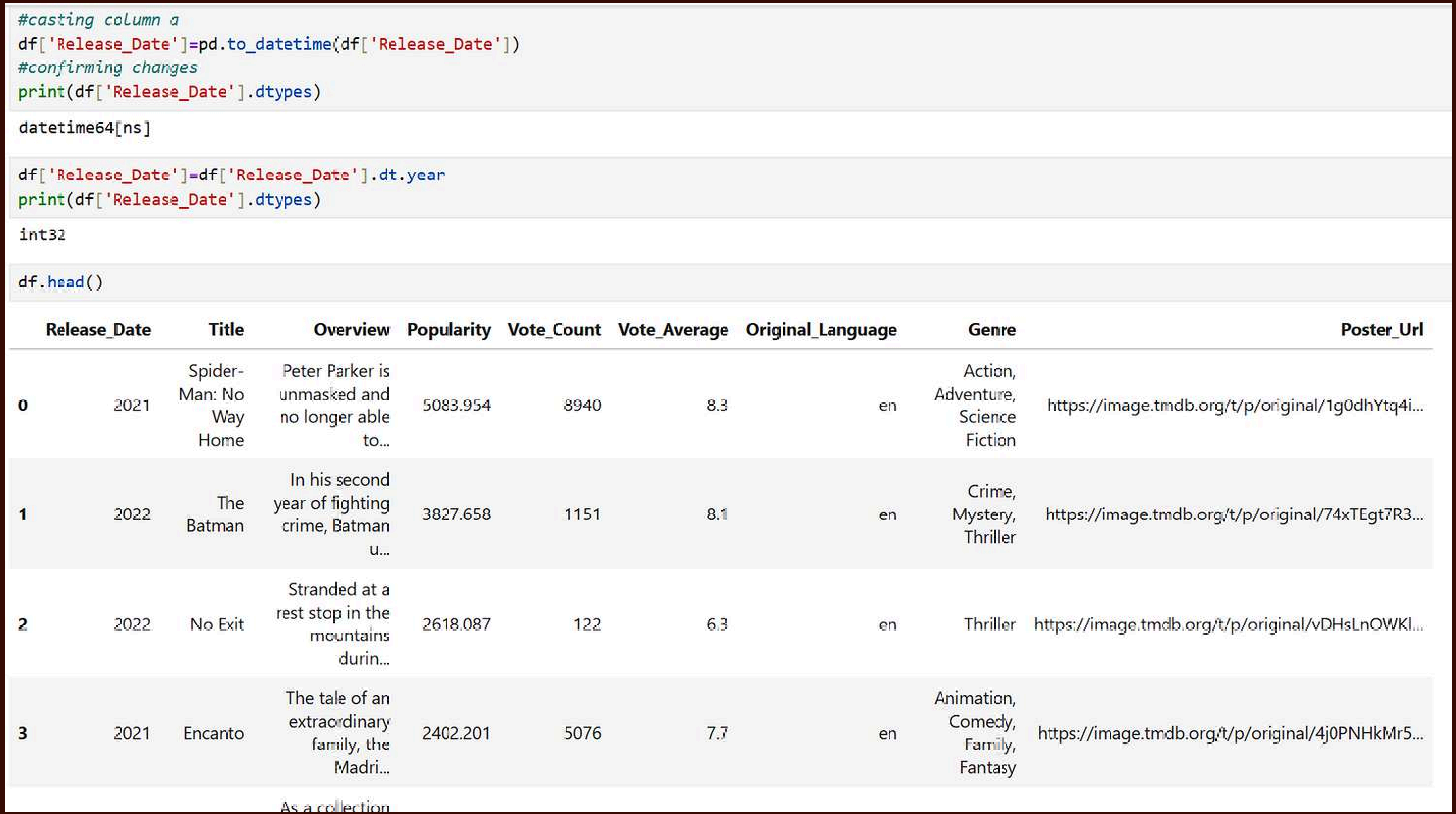


```
#importing libraries
import pandas as pd
import numpy as np
import seaborn as sns
import matplotlib.pyplot as plt


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



	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/1g0dhYtq4i...
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.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/4j0PNHkMr5...
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.tmdb.org/t/p/original/aq4Pwv5Xeu...

DATA PREPROCESSING



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

```
datetime64[ns]
```

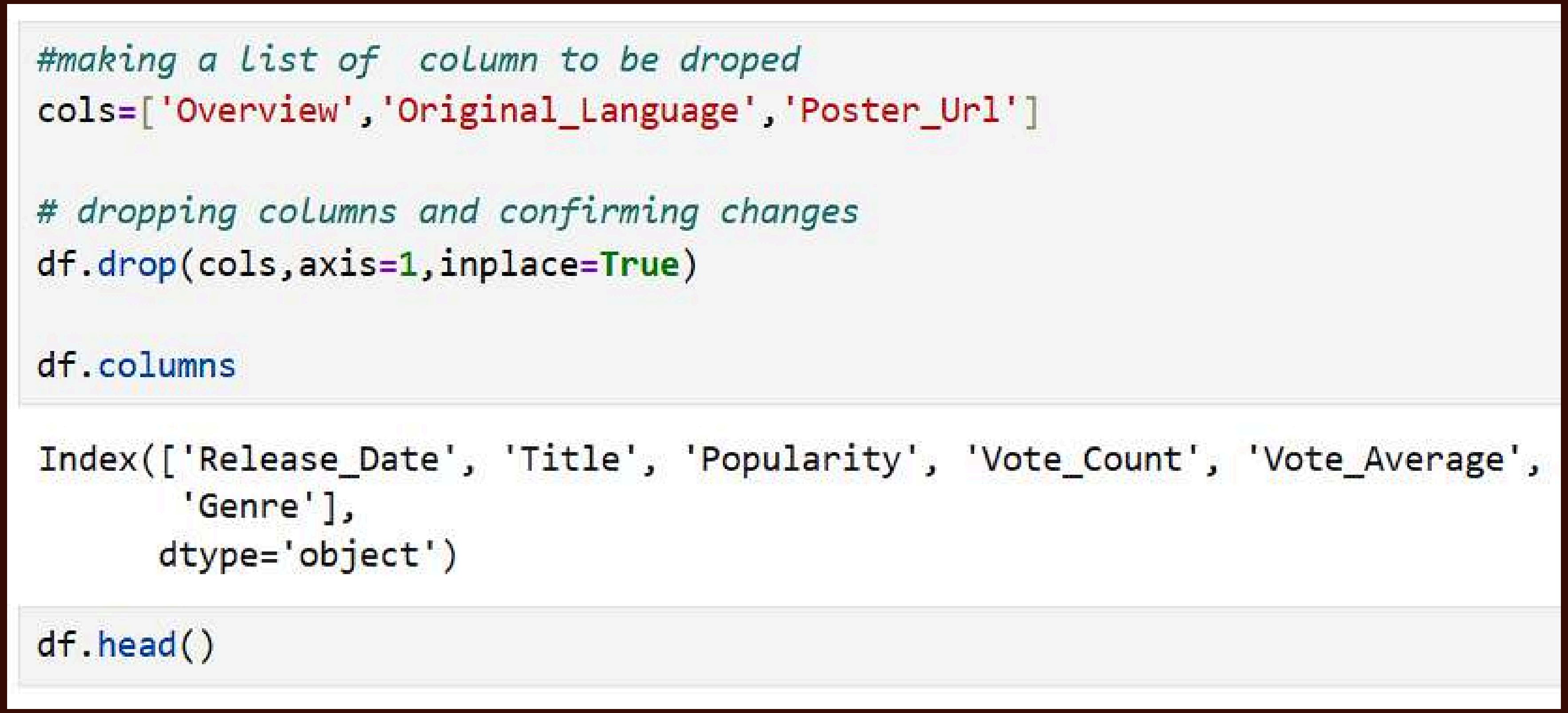
```
df['Release_Date']=df['Release_Date'].dt.year
print(df['Release_Date'].dtypes)
```

```
int32
```

```
df.head()
```

	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/1g0dhYtq4i...
1	2022	The Batman	In his second year of fighting crime, Batman u...	3827.658	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/4j0PNHkMr5...
			As a collection						

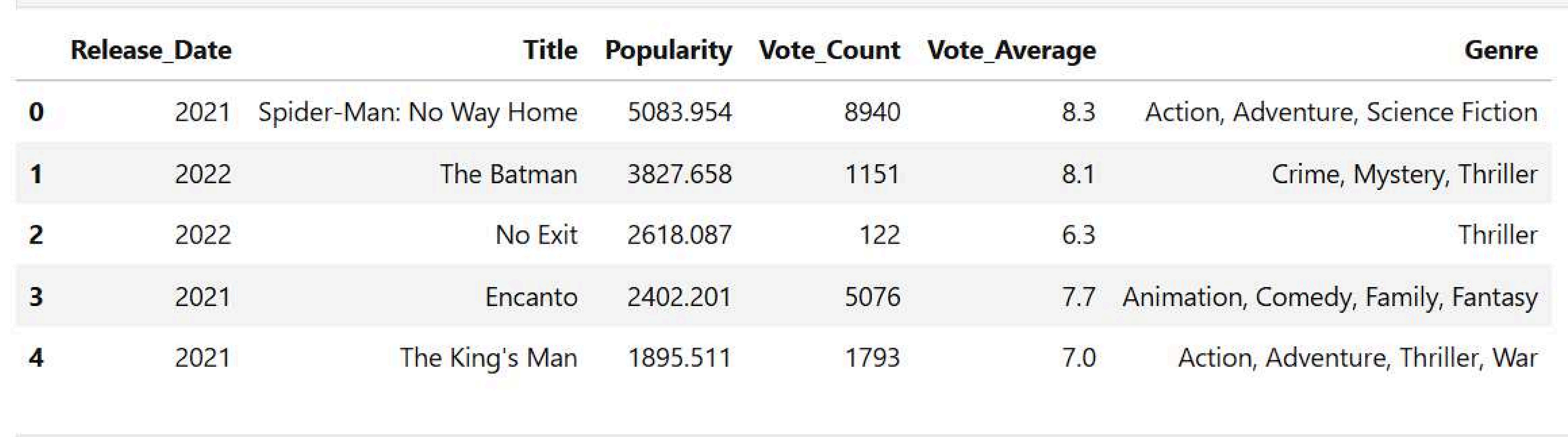
DATA PREPROCESSING



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

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

df.columns
```

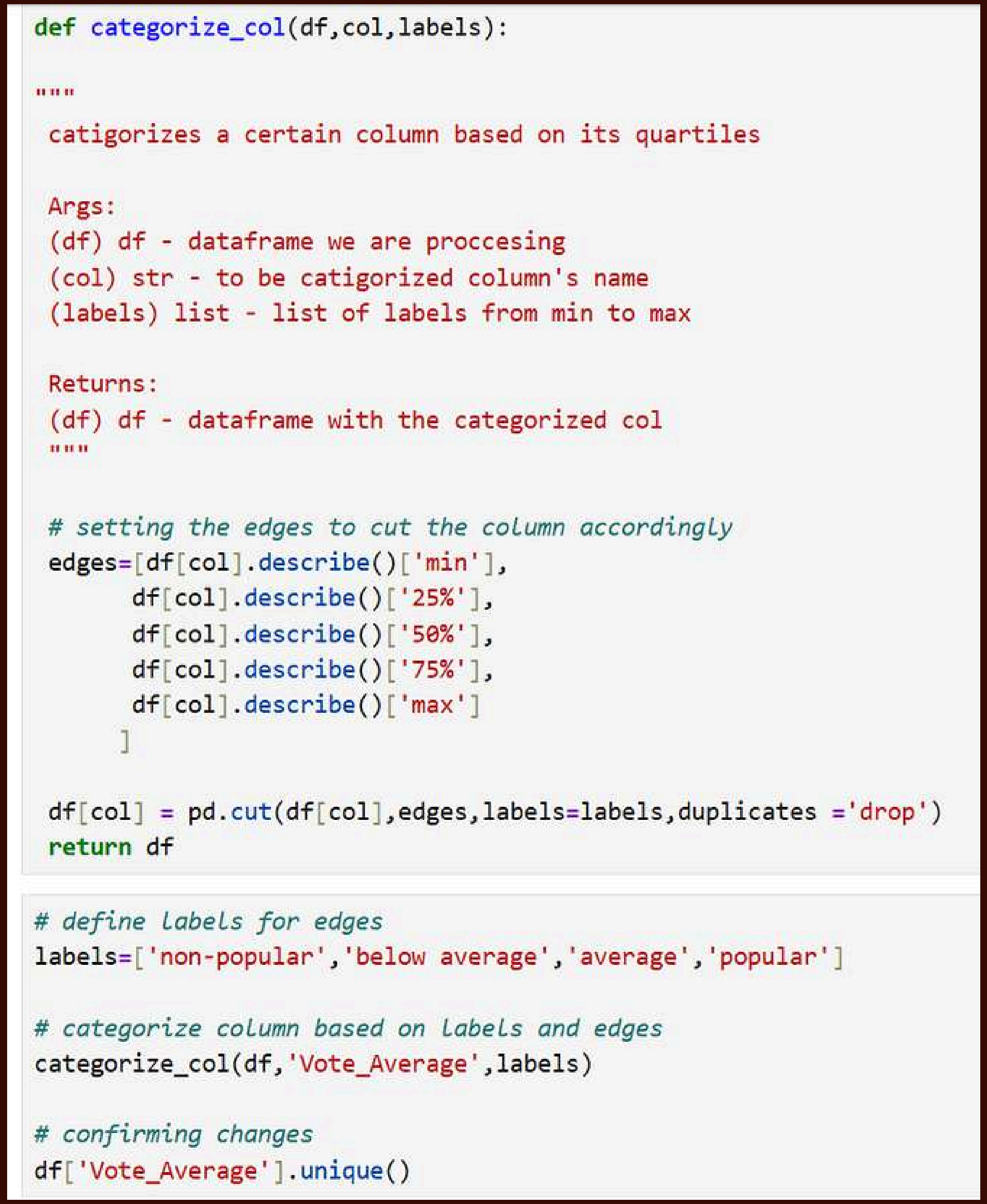
```
Index(['Release_Date', 'Title', 'Popularity', 'Vote_Count', 'Vote_Average',
       'Genre'],
      dtype='object')
```

```
df.head()
```



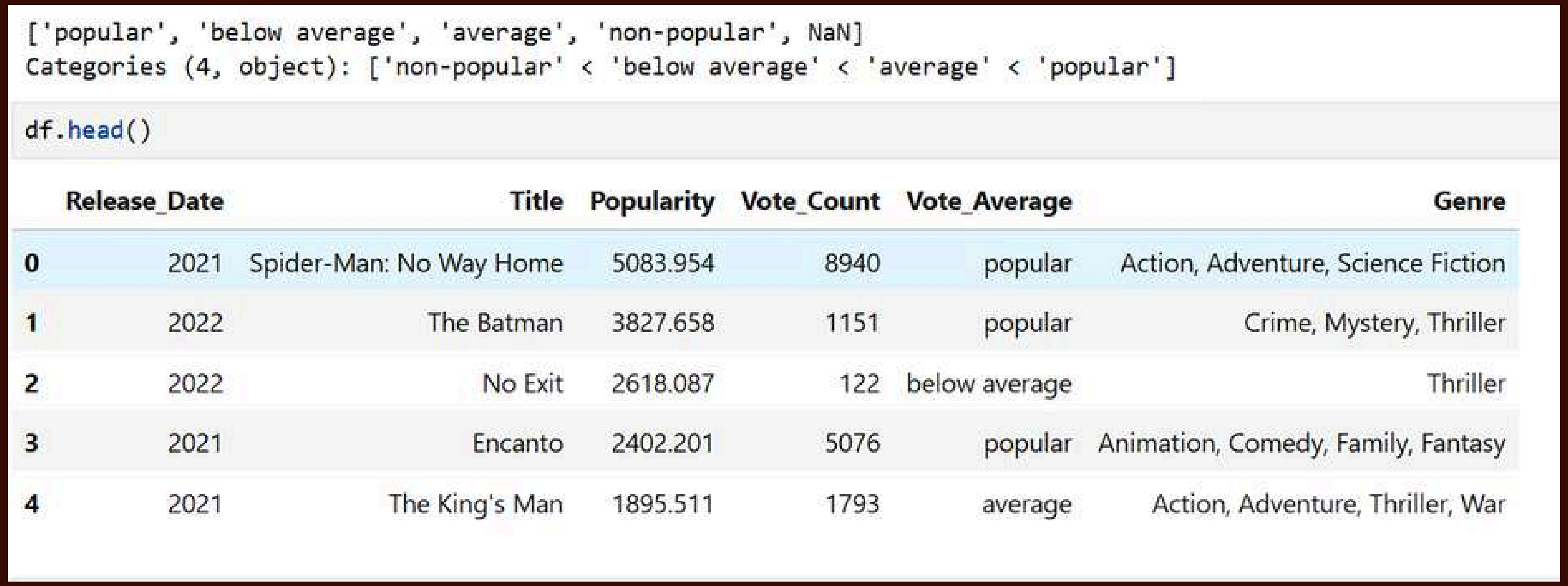
	Release_Date	Title	Popularity	Vote_Count	Vote_Average	Genre
0	2021	Spider-Man: No Way Home	5083.954	8940	8.3	Action, Adventure, Science Fiction
1	2022	The Batman	3827.658	1151	8.1	Crime, Mystery, Thriller
2	2022	No Exit	2618.087	122	6.3	Thriller
3	2021	Encanto	2402.201	5076	7.7	Animation, Comedy, Family,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

DATA PREPROCESSING



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

"""
 catigorizes a certain column based on its quartiles

 Args:
 (df) df - dataframe we are proccesing
 (col) str - to be catigorized column's name
 (labels) list - list of labels from min to max

 Returns:
 (df) df - dataframe with the categorized col
 """

 # 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=['non-popular','below average','average','popular']

# categorize column based on labels and edges
categorize_col(df,'Vote_Average',labels)

# confirming changes
df['Vote_Average'].unique()
```

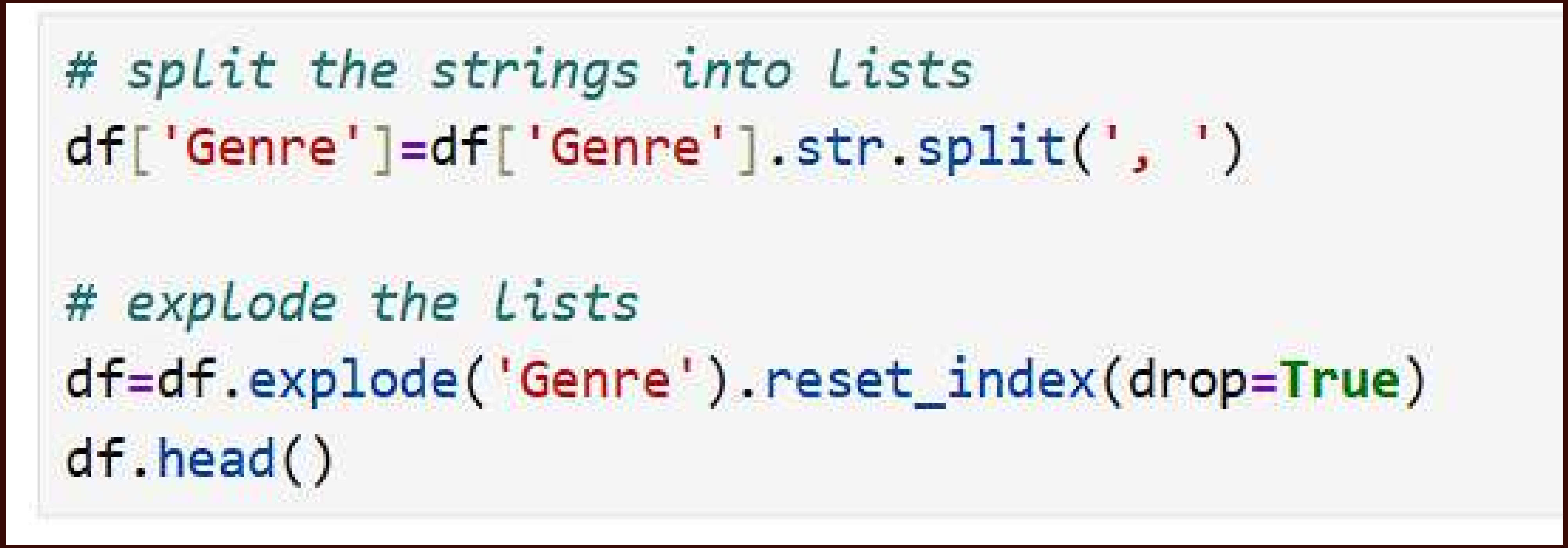


```
['popular', 'below average', 'average', 'non-popular', NaN]
Categories (4, object): ['non-popular' < 'below average' < 'average' < 'popular']
```

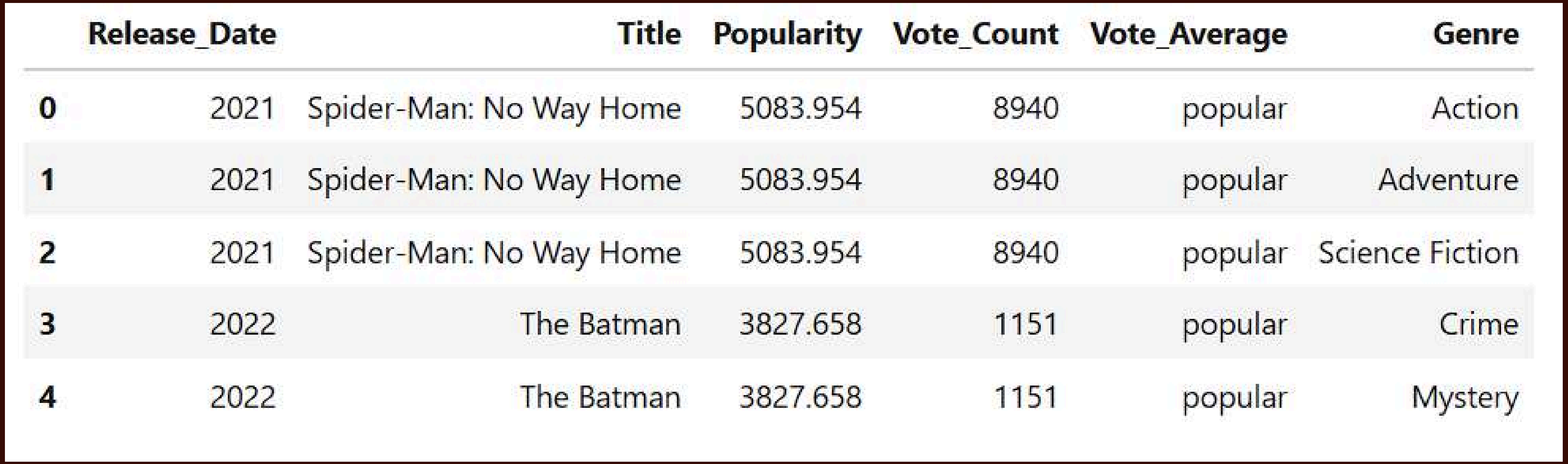
```
df.head()
```

	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	3827.658	1151	popular	Crime, Mystery, Thriller
2	2022	No Exit	2618.087	122	below average	Thriller
3	2021	Encanto	2402.201	5076	popular	Animation, Comedy, Family, Fantasy
4	2021	The King's Man	1895.511	1793	average	Action, Adventure, Thriller, War

DATA PREPROCESSING



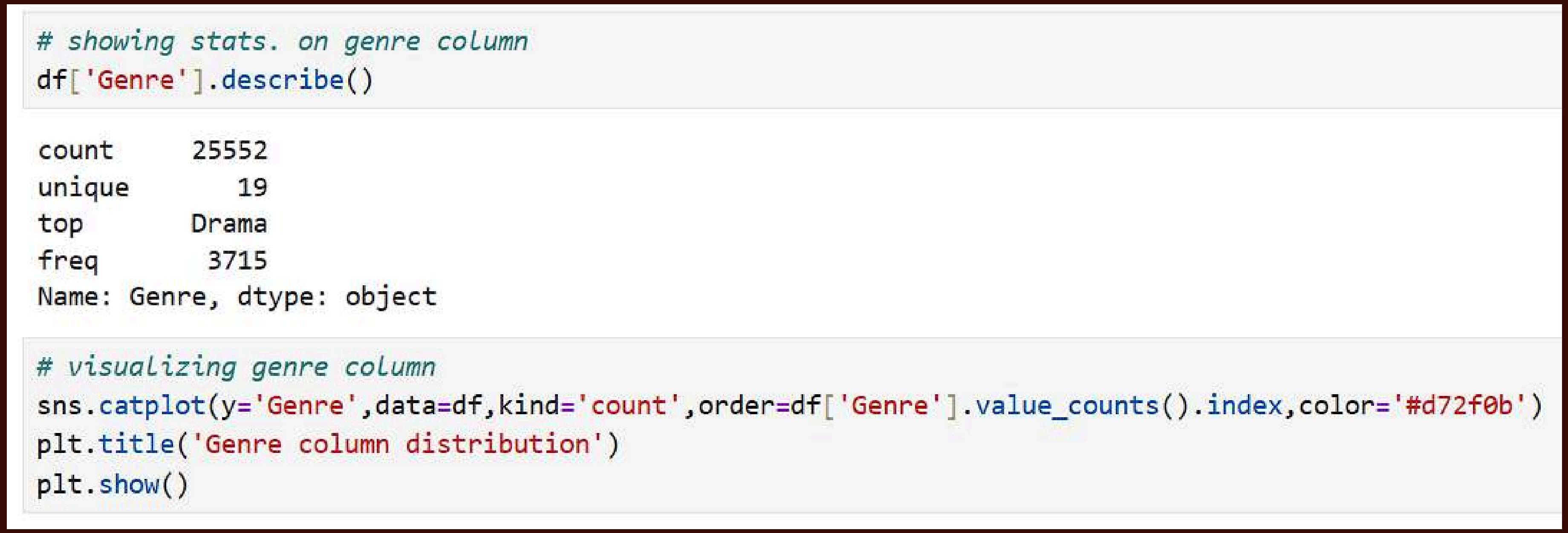
```
# split the strings into lists
df['Genre']=df['Genre'].str.split(', ')

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



	Release_Date	Title	Popularity	Vote_Count	Vote_Average	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	3827.658	1151	popular	Crime
4	2022	The Batman	3827.658	1151	popular	Mystery

DATA VISUALISATION

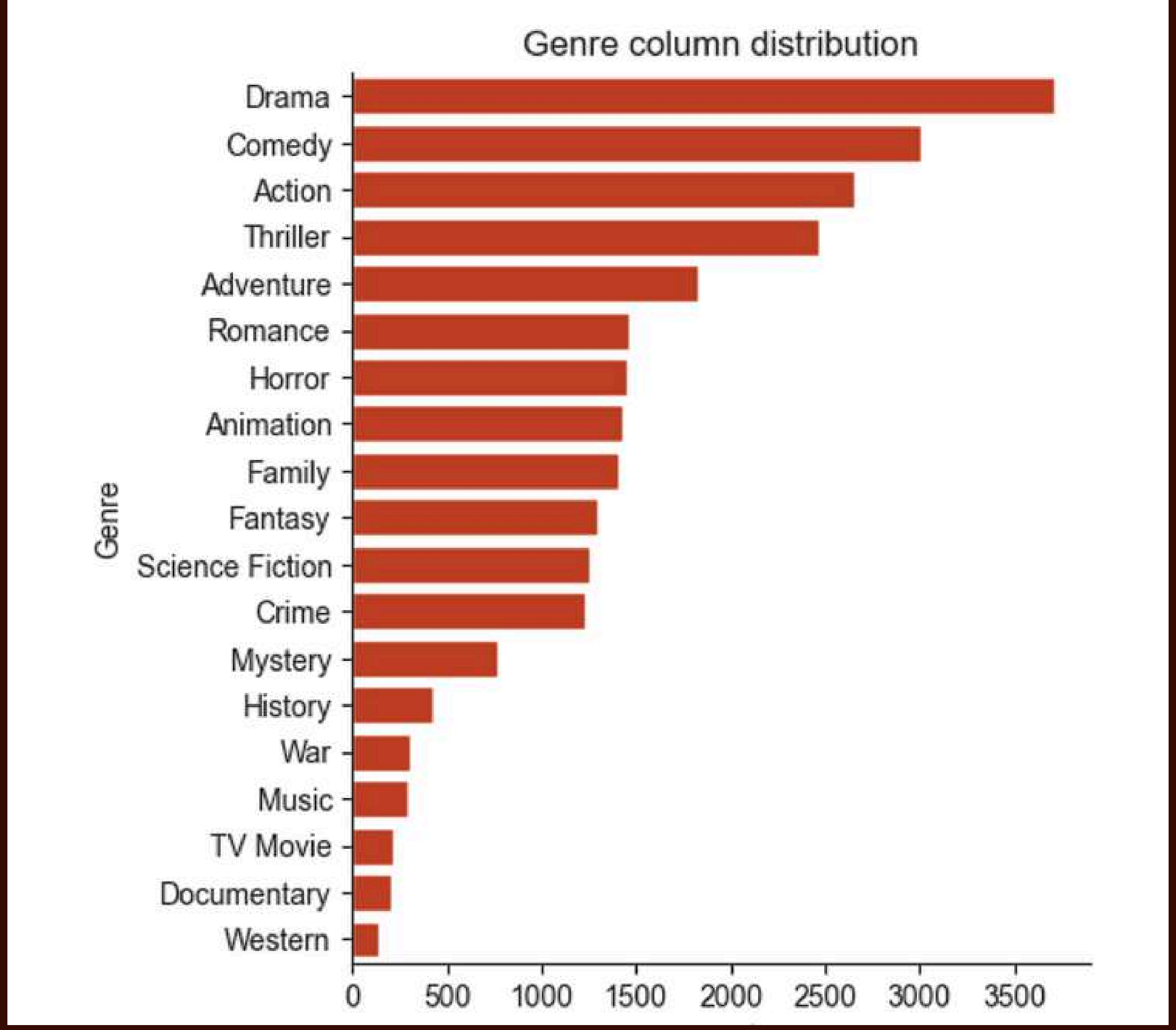
1)WHAT IS THE MOST FREQUENT GENRE OF MOVIES RELEASED ON NETFLIX?



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

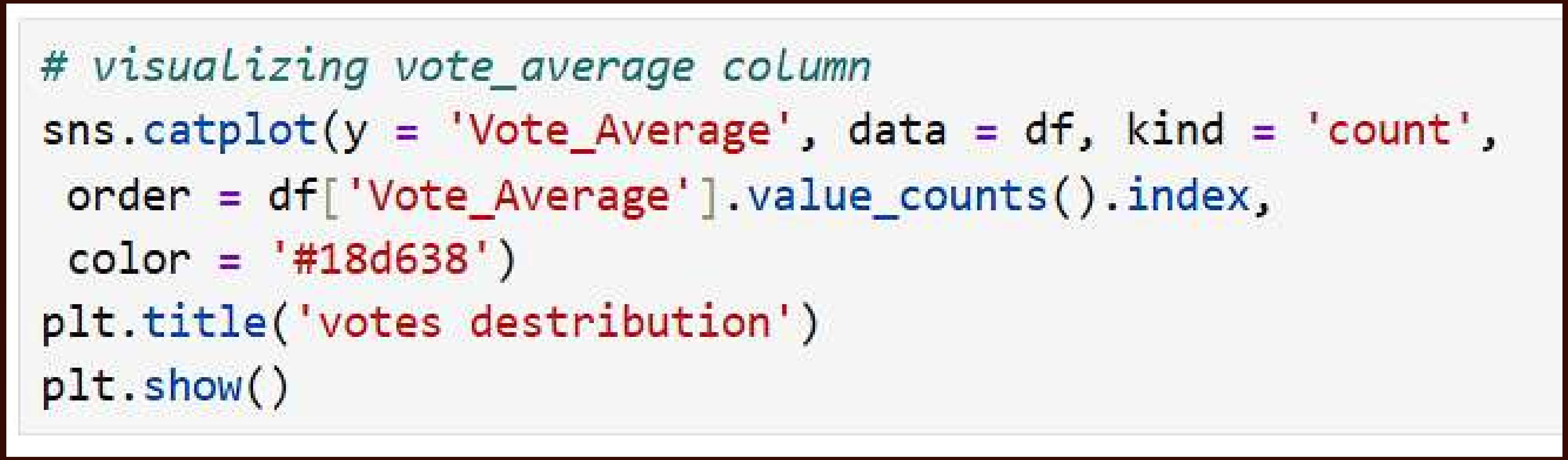
```
count      25552
unique        19
top        Drama
freq        3715
Name: Genre, dtype: object
```

```
# visualizing genre column
sns.catplot(y='Genre',data=df,kind='count',order=df['Genre'].value_counts().index,color='#d72f0b')
plt.title('Genre column distribution')
plt.show()
```

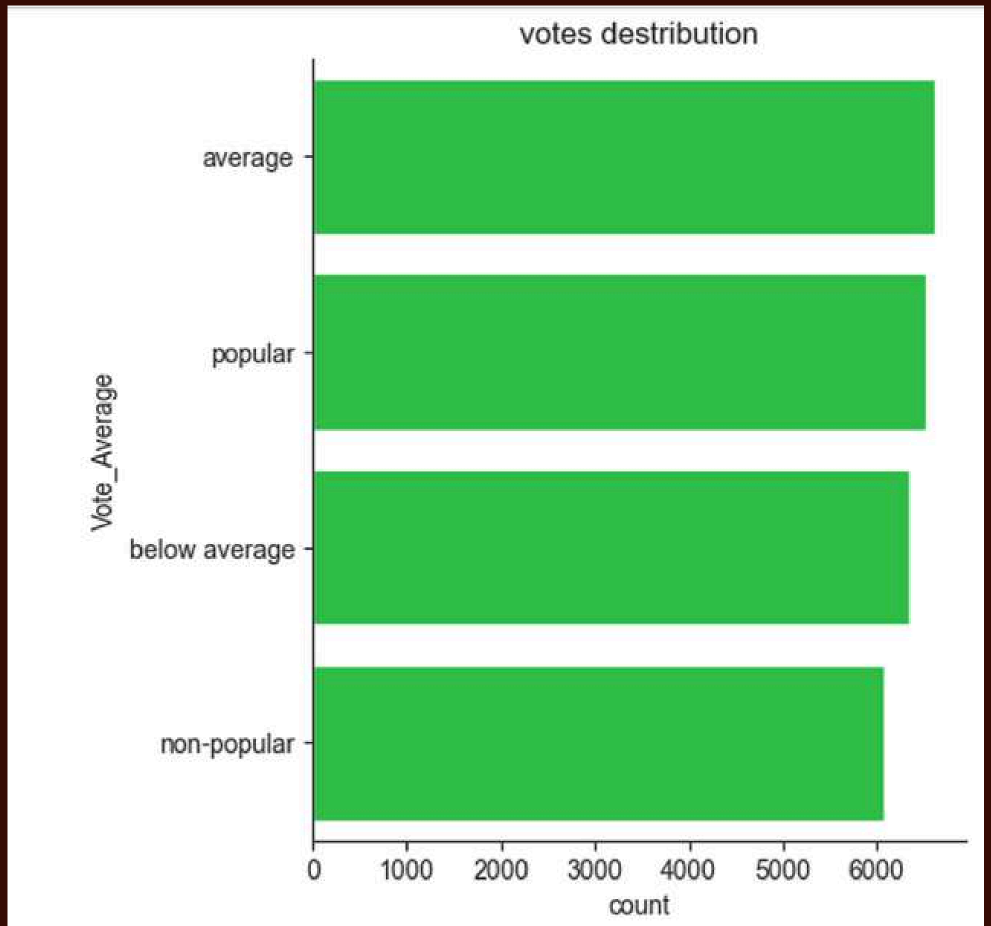


DATA VISUALISATION

2)WHICH HAS HIGHEST VOTES IN VOTE AVG COLUMN?

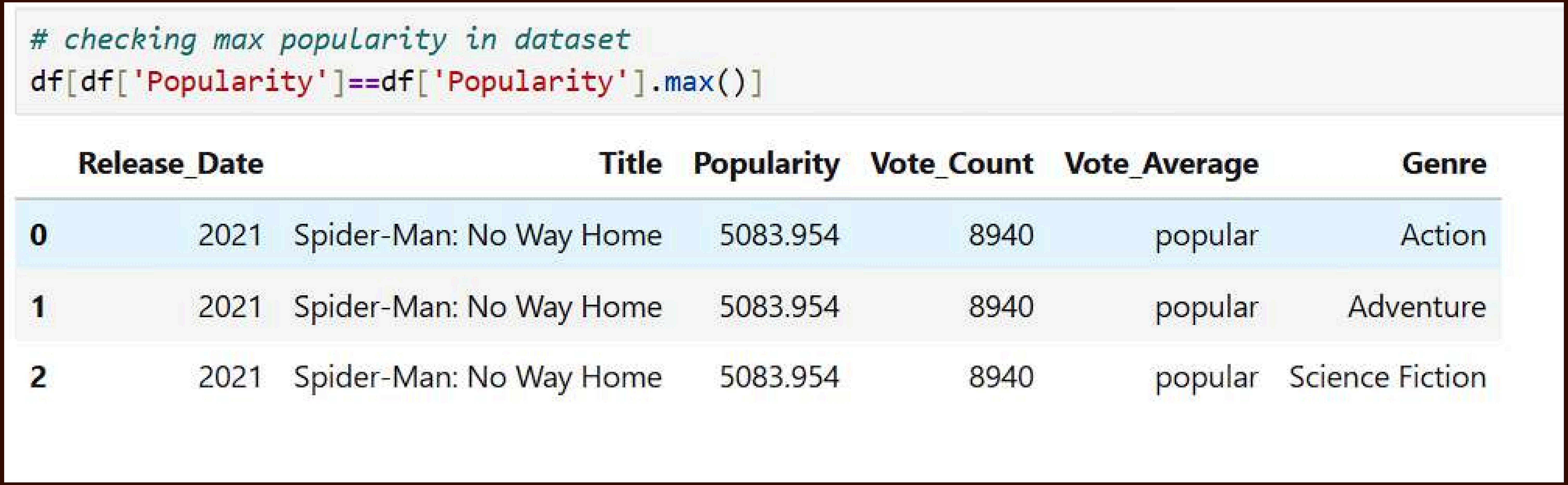


```
# visualizing vote_average column
sns.catplot(y = 'Vote_Average', data = df, kind = 'count',
 order = df['Vote_Average'].value_counts().index,
 color = '#18d638')
plt.title('votes destribution')
plt.show()
```



DATA VISUALISATION

3)WHAT MOVIE GOT THE HIGHEST POPULARITY? WHAT'S ITS GENRE?

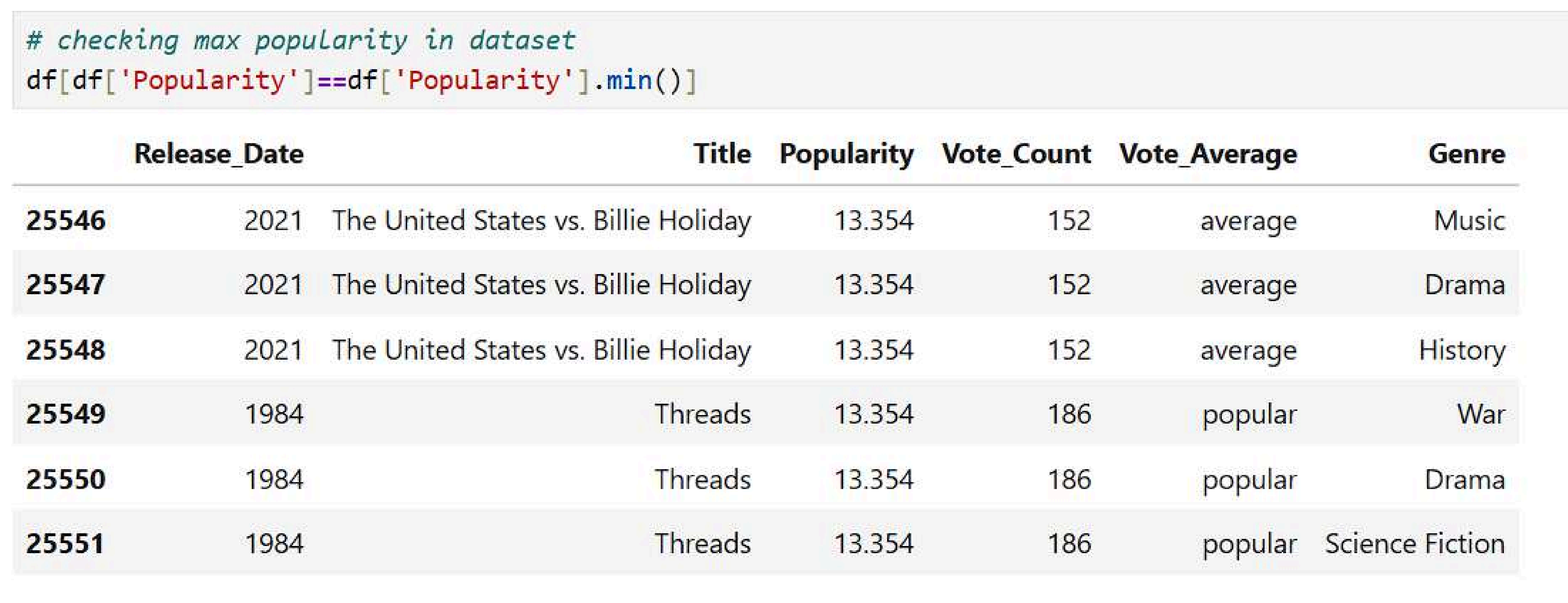


```python
# checking max popularity in dataset
df[df['Popularity']==df['Popularity'].max()]
```

	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

DATA VISUALISATION

4)WHAT MOVIE GOT THE LOWEST POPULARITY? WHAT'S ITS GENRE?

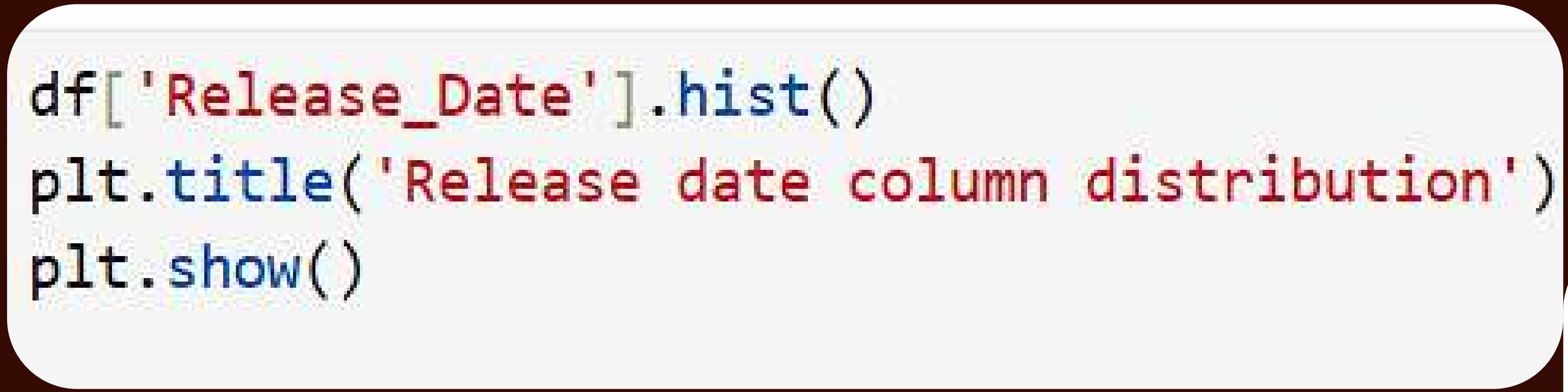


```python
# checking max popularity in dataset
df[df['Popularity']==df['Popularity'].min()]
```

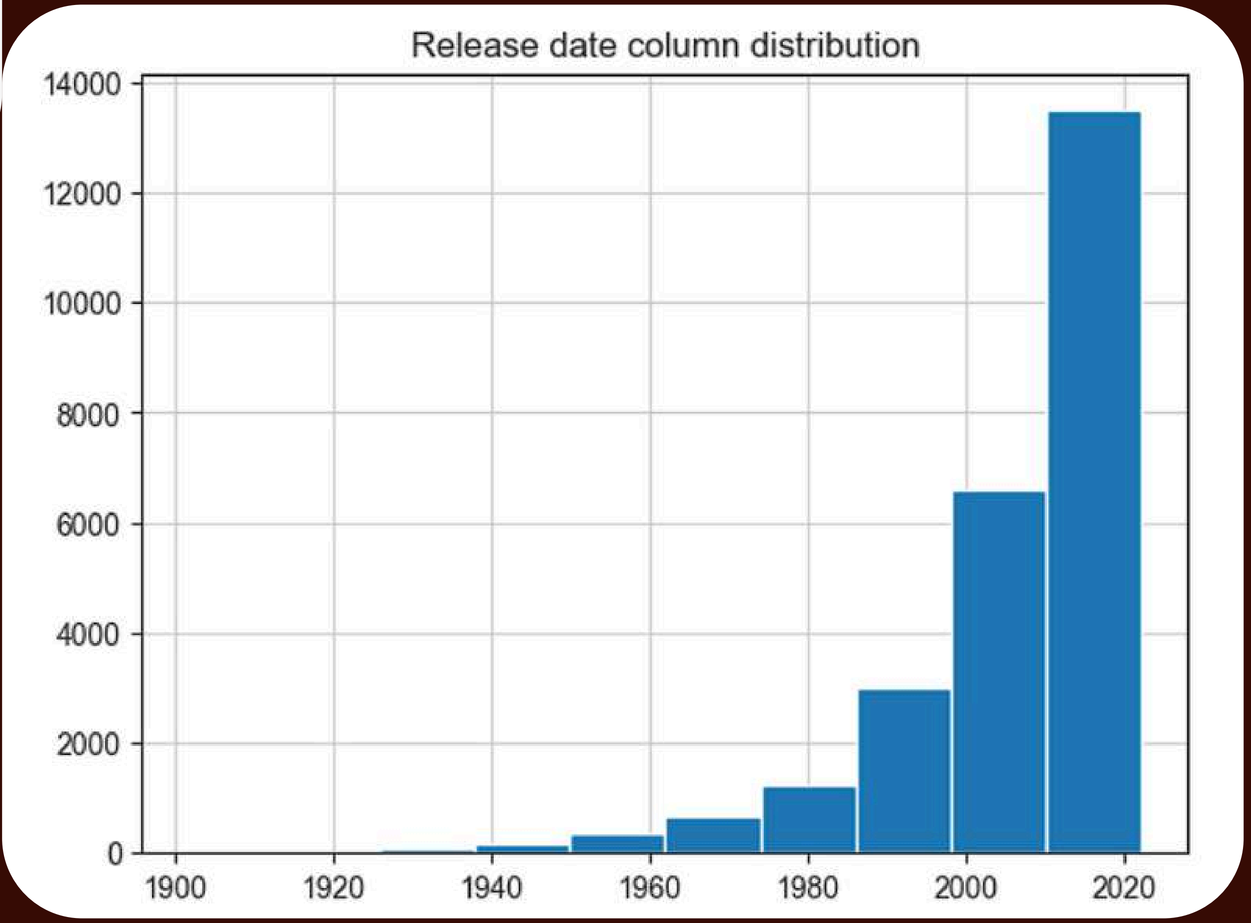
	Release_Date	Title	Popularity	Vote_Count	Vote_Average	Genre
25546	2021	The United States vs. Billie Holiday	13.354	152	average	Music
25547	2021	The United States vs. Billie Holiday	13.354	152	average	Drama
25548	2021	The United States vs. Billie Holiday	13.354	152	average	History
25549	1984	Threads	13.354	186	popular	War
25550	1984	Threads	13.354	186	popular	Drama
25551	1984	Threads	13.354	186	popular	Science Fiction

DATA VISUALISATION

4)WHAT MOVIE GOT THE LOWEST POPULARITY? WHAT'S ITS GENRE?



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

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`.

Q5: Which year has the most filmmed movies?

year 2020 has the highest filmming rate in our dataset.

THANK YOU

SEE YOU NEXT TIME