SVKM's NMIMS

Mukesh Patel School of Technology, Management and Engineering

Department of Computer Engineering

BTI, Integrated Computer VII Semester B Division

2023-2024

Subject: Data Extraction and Processing

Project Report

Group No: 1

Team Member's Details:

Roll No's	Names
C034	Reneeka Nadkarni
C046	Chaitanya Ajgaonkar
C053	Milberg Noronha
C065	Kirtan Thakkar

Project Title: Exploratory Data Analysis on Movie Review website - Rotten Tomatoes

Dataset Title: Rotten Tomatoes Movies Dataset

URL for Data Set Download: https://www.kaggle.com/datasets/stefanoleone992/rotten-tomatoes-movies-and-critic-reviews-dataset/

Data Set Description in Brief: Rotten Tomatoes website allows to compare the ratings given by regular users (audience score) and the ratings given (tomatometer) by certified members of various writing guilds or film critic-associations. In the dataset, each record represents a movie available on Rotten Tomatoes, with the URL used for the scraping, movie title, description, genres, duration, director, actors, users' ratings, and critics' ratings.

<u>Data set Exploration code/screenshot and your inference:</u>

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

```
[2] df=pd.read_csv("/content/rotten_tomatoes_movies.csv")
    df.head(10)
```

	rotten_tomatoes_link	movie_title	movie_info	critics_consensus	content_rating	genres	directors	authors	actors	original_release_date	...	production_company	tomatometer_status	tomato
0	m/0814255	Percy Jackson & the Olympians: The Lightning T...	Always trouble-prone, the life of teenager Per...	Though it may seem like just another Harry Pot...	PG	Action & Adventure, Comedy, Drama, Science Fic...	Chris Columbus	Craig Titley, Chris Columbus, Rick Riordan	Logan Lerman, Brandon T. Jackson, Alexandra Da...	2010-02-12	...	20th Century Fox	Rotten	
1	m/0878835	Please Give	Kate (Catherine Keener) and her husband Alex (...	Nicole Holofcener's newest might seem slight i...	R	Comedy	Nicole Holofcener	Nicole Holofcener	Catherine Keener, Amanda Peet, Oliver Platt, R...	2010-04-30	...	Sony Pictures Classics	Certified-Fresh	
2	m/10	10	A successful, middle-aged Hollywood songwriter...	Blake Edwards' bawdy comedy may not score a pe...	R	Comedy, Romance	Blake Edwards	Blake Edwards	Dudley Moore, Bo Derek, Julie Andrews, Robert ...	1979-10-05	...	Waner Bros.	Fresh	
3	m/1000013-12_angry_men	12 Angry Men (Twelve Angry Men)	Following the closing arguments in a murder tr...	Sidney Lumet's feature debut is a superbly wri...	NR	Classics, Drama	Sidney Lumet	Reginald Rose	Martin Balsam, John Fiedler, Lee J. Cobb, E.G....	1957-04-13	...	Criterion Collection	Certified-Fresh	
4	m/1000079-20000_leagues_under_the_sea	20 000 Leagues Under The Sea	In 1866, Professor Pierre M. Aronnax (Paul Luk...	One of Disney's finest live-action adventures,...	G	Action & Adventure, Drama, Kids & Family	Richard Fleischer	Earl Felton	James Mason, Kirk Douglas, Paul Lukas, Peter L...	1954-01-01	...	Disney	Fresh	

Imported the required libraries and the dataset. Found out the head of the data set.

```
df.isnull().sum()
```

```
rotten_tomatoes_link                0
movie_title                         0
movie_info                        321
critics_consensus                8578
content_rating                      0
genres                             19
directors                         194
authors                          1542
actors                            352
original_release_date            1166
streaming_release_date            384
runtime                           314
production_company                499
tomatometer_status                 44
tomatometer_rating                 44
tomatometer_count                  44
audience_status                   448
audience_rating                   296
audience_count                    297
tomatometer_top_critics_count       0
tomatometer_fresh_critics_count     0
tomatometer_rotten_critics_count    0
dtype: int64
```

Discovered number of null values in each column of the dataset.

```
[3] df.shape
```

```
(17712, 22)
```

Found out the shape of the dataset, it contains 17712 rows and 22 columns.

```
df.dtypes
```

```
rotten_tomatoes_link                object
movie_title                         object
movie_info                          object
critics_consensus                   object
content_rating                      object
genres                              object
directors                           object
authors                             object
actors                              object
original_release_date                object
streaming_release_date               object
runtime                             float64
production_company                   object
tomatometer_status                   object
tomatometer_rating                  float64
tomatometer_count                   float64
audience_status                      object
audience_rating                     float64
audience_count                      float64
tomatometer_top_critics_count         int64
tomatometer_fresh_critics_count       int64
tomatometer_rotten_critics_count      int64
dtype: object
```

Checked all the datatypes of the attributes of the dataset.

Data set Preprocessing code/screenshot and your inference

```
df.dropna(subset=["audience_rating"],inplace=True)
df.dropna(subset=["runtime"],inplace=True)
df.dropna(subset=["original_release_date"],inplace=True)
df.dropna(subset=["genres"],inplace=True)
df.dropna(subset=["tomatometer_rating"],inplace=True)
df.dropna(subset=["production_company"],inplace=True)

df.drop('critics_consensus', axis=1, inplace=True)

#Removing null values and unwanted coloumns
```

```
[7] df.shape

(15930, 21)
```

Dropped the rows which contained null values for the above attributes. Also removed the ‘crictics_consensus’ column since it was not required in our analysis. We found the shape of the updated data set. Now it contains 15930 rows and 21 columns.

```
df.isnull().sum()

rotten_tomatoes_link                 0
movie_title                          0
movie_info                          15
content_rating                       0
genres                               0
directors                          133
authors                           1142
actors                             226
original_release_date                0
streaming_release_date              92
runtime                              0
production_company                   0
tomatometer_status                   0
tomatometer_rating                   0
tomatometer_count                    0
audience_status                    139
audience_rating                      0
audience_count                       1
tomatometer_top_critics_count        0
tomatometer_fresh_critics_count      0
tomatometer_rotten_critics_count     0
dtype: int64
```

Checked if the null values have been removed for the required attributes.

```
[21] df["audience_rating"]=df["audience_rating"].astype("int64")
     df["tomatometer_rating"]=df["tomatometer_rating"].astype("int64")
     df["original_release_date"]=df["original_release_date"].astype("datetime64")

     #Converting data formats
```

```
df.dtypes
```

```
rotten_tomatoes_link                      object
movie_title                               object
movie_info                                object
content_rating                            object
genres                                    object
directors                                 object
authors                                   object
actors                                    object
original_release_date               datetime64[ns]
streaming_release_date                    object
runtime                                  float64
production_company                        object
tomatometer_status                        object
tomatometer_rating                         int64
tomatometer_count                        float64
audience_status                           object
audience_rating                            int64
audience_count                           float64
tomatometer_top_critics_count              int64
tomatometer_fresh_critics_count            int64
tomatometer_rotten_critics_count           int64
dtype: object
```

Converted the data types for ‘audience_rating’ and ‘tomatometer_rating’ to int64 and ‘original_release_date’ to ‘datetime64’

```python
print(df["audience_rating"].unique())
print(df["tomatometer_rating"].unique())
print(df["runtime"].unique())
print(df["original_release_date"].unique())
print(df["content_rating"].unique())
#Determining whether data has random data in it
```

```
[ 53  64  97  74  37  86  79  87  40  66  35  57  82  80  89  75  60  91
  63  71  61  30  48  56  83  39  88  65  73  85  42  76  78  67  94  20
  68  58  55  84  93  45  29  24  32  43  59  69  33  52  22  46  34  50
  54  36  21  27  62  72  44  18  47  26  90  38  81  28  41  19  17  49
  11  92  31  12  70  77  25  95  51  23  16  14  15  96   9  10 100   5
   8  13  99   7   0  98   6   4]
[ 49  87  67 100  89   8  96  20  82  80  91  25   4  69  92  75   0  31
  63  41  93  22  33  21  15  64  32  61  52  24  14  44  60  97  36  99
  42  30  10  85  84  95  88  81  86  78  54  35  83  58  17  40  55  50
  45  13  73  37  47   7  56  34  38  26  76  16  12  94  19  70  68   9
  62  72  71  29  79   5  53  46  39  28  23  74  48  11  27  90  43  77
  51  98  59  57  65  66   6  18   3   2   1]
[119.  90. 122.  95. 127. 109.  80.  92. 103.  97. 100. 110. 143. 101.
 124.  87.  86. 102.  83.  77. 126.  94. 121.  93.  88.  75. 108. 150.
  71. 107. 130.  98.  91. 192.  96. 141. 114. 104. 137. 138.  85.  76.
 116. 106.  81.  84.  99.  89. 120.  73. 115. 146. 147. 165. 151.  78.
 113. 128. 118. 140. 144. 105. 117. 111. 112. 145.  82. 123. 125.  62.
  79. 133. 170. 129. 156.  66.  58. 132.  74. 179. 131. 135.  69. 243.
  65. 197. 207. 169.  22. 149. 153. 181. 139. 134. 178.  72. 160.  70.
 152. 174. 183. 171.   8. 184. 176. 220. 142. 189. 136. 201. 157. 163.
 154. 254. 155.  60. 191. 172. 177. 242. 180. 158. 187.  39.  46.  47.
  40. 168.  35. 240. 164.  68. 175. 162. 195. 148. 166.  67. 167. 203.
 188. 161. 190.  63. 159. 173.  45.  64.  44.  53. 200. 223. 222.  43.
  13. 218. 210.  56.  42. 213. 193. 209. 196. 216.  50.  61. 266.  49.
  32. 250.  52.  55.  59. 238. 206. 182.  57. 208.  48.  28.  15.  41.]
['2010-02-12T00:00:00.000000000' '2010-04-30T00:00:00.000000000'
 '1979-10-05T00:00:00.000000000' ... '1981-10-02T00:00:00.000000000'
 '1964-12-17T00:00:00.000000000' '1964-06-17T00:00:00.000000000']
['PG' 'R' 'NR' 'G' 'PG-13' 'NC17']
```

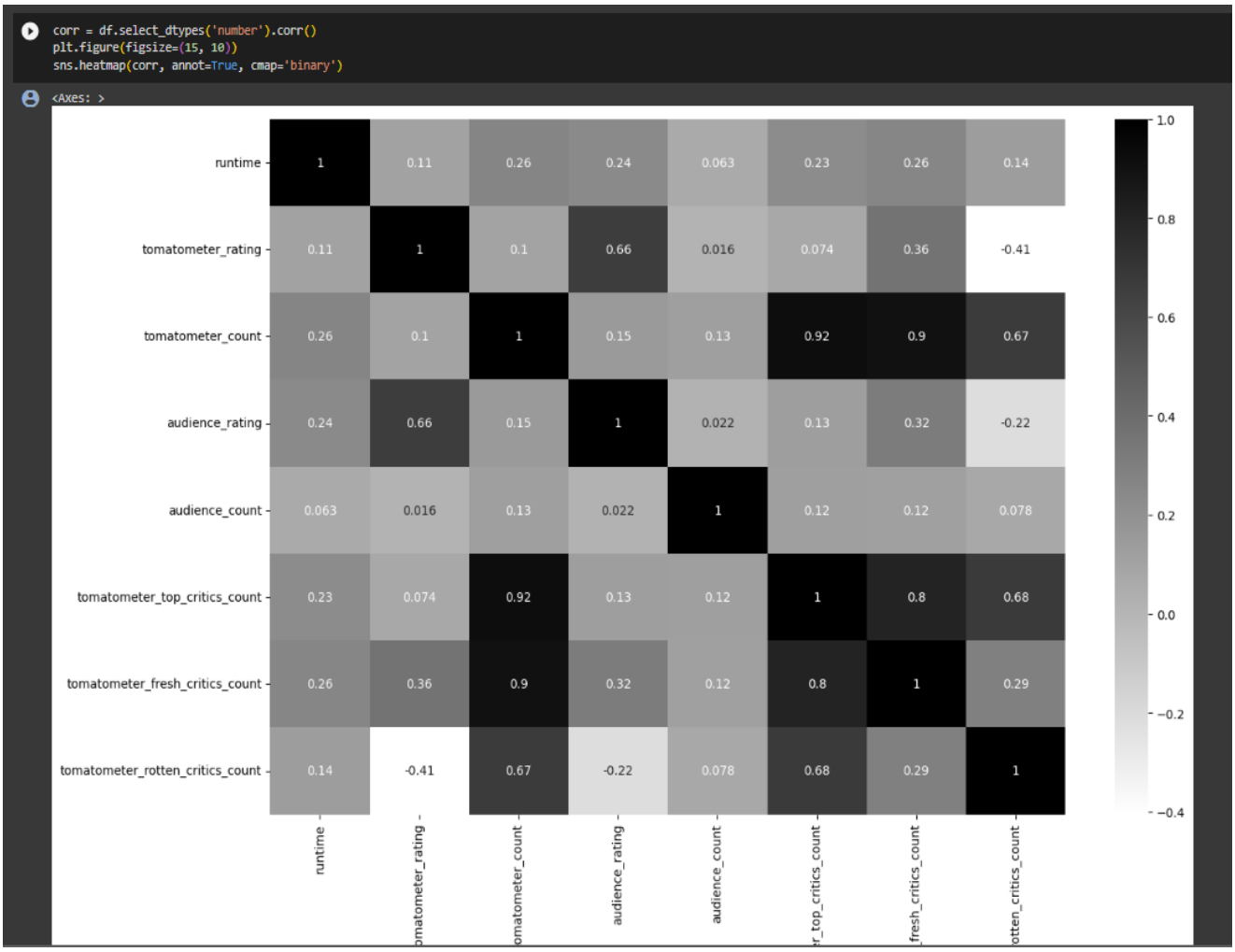
Checked the unique values of the above attributes.

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

```
0
```

Checked for duplicate values in the data set.

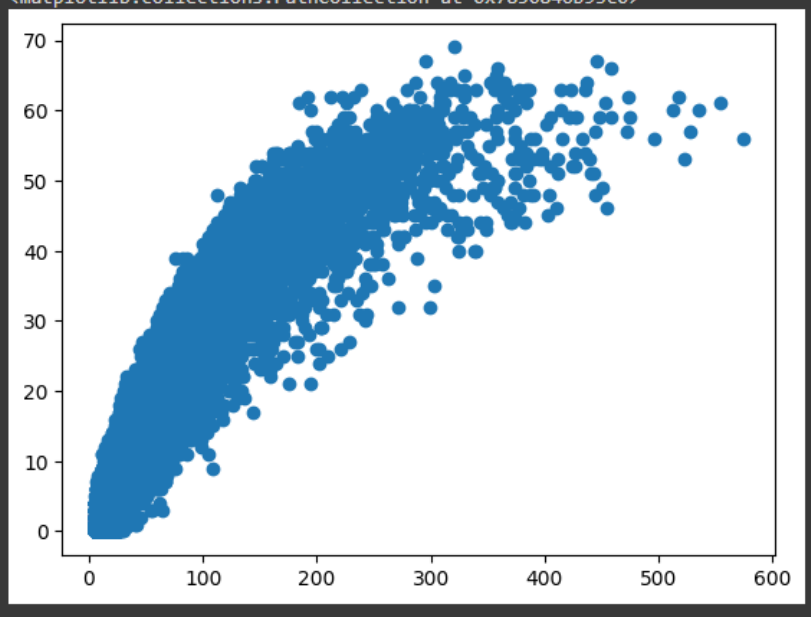
Data set visualization/screenshot and your inference:



Created a heatmap to find correlation among all the attributes. From the heatmap, we saw that ‘tomatometer_count’ and ‘tomatometer_top_critcis_count’ are highly positively correlated (0.92). We also found a few negatively correlated attributes.

```
plt.scatter(df['tomatometer_count'],df['tomatometer_top_critics_count'])
```

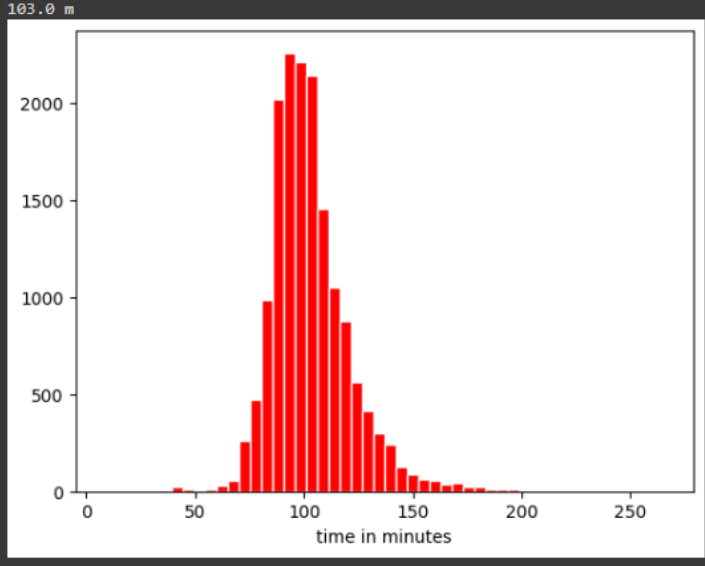
```
<matplotlib.collections.PathCollection at 0x7850846b93c0>
```



Plotted a scatter plot for ‘tomatometer_count’ and ‘tomatometer_top_critcis_count’. They are highly positively correlated (0.92).

```
plt.hist(x=df["runtime"],bins=50,edgecolor="white",color="red")
plt.xlabel("time in minutes")
print("Average time of movies")
print(df["runtime"].mean().round(),"m")
```

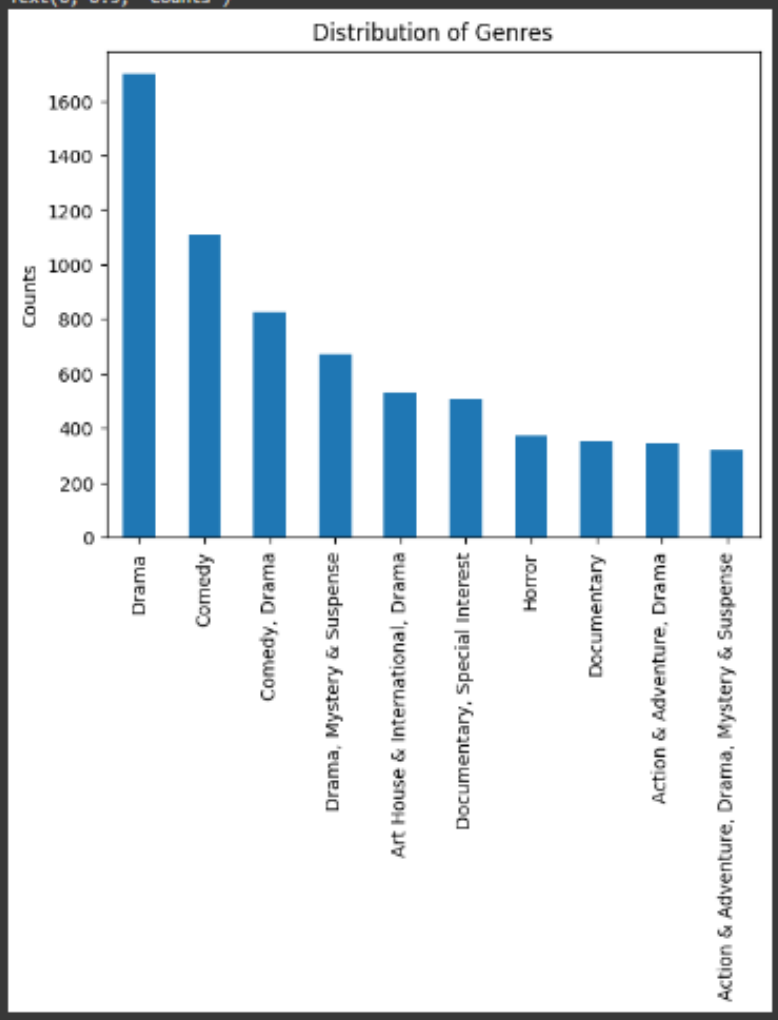
```
Average time of movies
103.0 m
```



Plotted a histogram for ‘runtime’ which is slightly right-skewed and found the average time for all the movies (103 minutes).

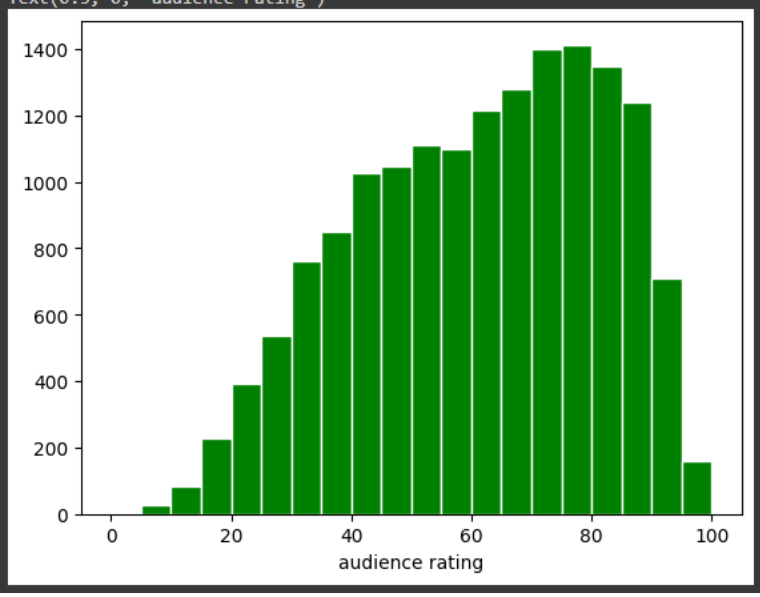
```
df['genres'].value_counts()[:10].plot(kind='bar')
plt.title('Distribution of Genres')
plt.ylabel('Counts')
```

```
Text(0, 0.5, 'Counts')
```



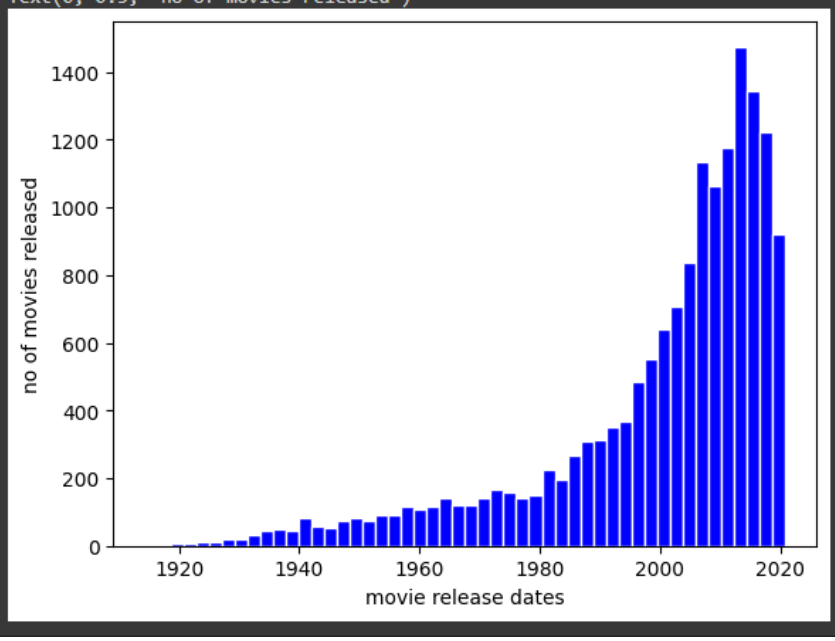
Found the top 10 genres and plotted the distribution for the same.

```
[34] plt.hist(df["audience_rating"],bins=20,edgecolor='white',color="green")
     plt.xlabel("audience rating")

     Text(0.5, 0, 'audience rating')
```



Plotted histogram for showing the distribution of ‘audience_rating’. The plot is left-skewed.

```
plt.hist(df["original_release_date"], bins=50,edgecolor="white",color="blue")
plt.xlabel("movie release dates")
plt.ylabel("no of movies released")

Text(0, 0.5, 'no of movies released')
```



Plotted a histogram to show the distribution of number of movies released throughout the years. The plot is left-skewed which indicates a greater number of movies were released in recent years.

```python
print("oldest movie:")
oldest=min(df["original_release_date"])
print(df[(df['original_release_date']==oldest)].iloc[:,[1,8]].to_string())

print("latest movie:")
latest=max(df["original_release_date"])
print(df[(df['original_release_date']==latest)].iloc[:,[1,8]].to_string())
```

```
oldest movie:
     movie_title original_release_date
4205     Cabiria            1914-06-01
latest movie:
        movie_title original_release_date
1621  Billy the Kid            2020-09-30
```

Found out the oldest and latest movie from the updated data set by extracting the movie name from column 1 (movie_title) and the release date from column 8 (original_release_date).

Dataset-2

```python
[43] print("Movies with a rating of 90% and over:")
limit=90
df2=pd.DataFrame({"Movie Name":df[(df['audience_rating']>=limit)].iloc[:,1],"Release date":df[(df['audience_rating']>=limit)].iloc[:,8],"Duration":df[(df['audience_rating']>=limit)].iloc[:,10
print(df2.to_string())
```

```
Movies with a rating of 90% and over:
                                                                                                Movie Name Release date  Duration Content Rating                        Production Com
3                                                                                 12 Angry Men (Twelve Angry Men)   1957-04-13      95.0             NR                      Criterion Collec
23                                                                                                     Bandwagon   1997-09-12     103.0              R                   Avalanche Entertain
50                                                                                                         Aliens   1986-07-18     137.0              R                        20th Century
52                                                                                                    All About Eve   1950-01-01     138.0             PG                        20th Century
62                                                                                                     Street Fight   2005-01-01      86.0             NR                Marshall Curry Product
155                                                                                                     The Hammer   2007-04-26      93.0              R                             IFC F
215                                                                                                Okuribito (Departures)   2009-05-29     130.0          PG-13                        Regent Relea
234                                                                                                   The Apartment   1960-06-15     125.0             NR                          United Art
306                                                                                     Beauty and The Beast (La Belle et la bête)   1946-01-01      95.0             NR                         Lopert Pict
314                                                                                                    The Big Sleep   1946-08-31     114.0             PG                    Warner Bros. Pict
339                                                                                                 Brief Encounter   1945-11-26      86.0             NR                       Universal Pict
346                                                                                       Butch Cassidy and the Sundance Kid   1969-10-24     110.0             PG                        20th Century
357                                                                                                       Casablanca   1942-11-26     102.0             PG                    Warner Bros. Pict
366                                                                                                          Charade   1963-12-05     114.0              G                     Madacy Entertain
420                                                                                                     East of Eden   1955-04-10     115.0             PG                    Warner Bros. Pict
424                                                                                                  The Elephant Man   1980-10-03     125.0             PG                              Param
461                                                                                                      The General   1927-02-05      83.0              R                   United Artists F
465                                                                                                            Glory   1989-12-15     122.0              R                                 Tri
492                                                                                 The Hidden Fortress (kakushi-toride No San-akunin)   1958-01-01     139.0             NR                  Media Home Entertain
513                                                                                                  Imitation of Life   1959-04-30     125.0             NR                        Universal Stu
520                                                                                                   Inherit the Wind   1960-11-01     127.0             PG                   MGM Home Entertain
552                                                                                                 The King of Comedy   1983-01-01     101.0             PG
557                                                                                                           Laura   1944-10-11      88.0             NR                        20th Century
562                                                                                                 The Lion in Winter   1968-10-30     132.0             PG                     Nelson Entertain
581                                                                                                              M   1931-08-31      99.0             NR                                 For
588                                                                                                 The Maltese Falcon   1941-10-18     100.0             PG                            Warner B
590                                                                                             The Manchurian Candidate   1962-10-24     126.0          PG-13                        MGM/UA Clas
600                                                                                                       Metropolis   1927-03-13     123.0          PG-13                      Paramount Pict
```

Created another data frame for movies with an audience rating of 90% and above.

Preprocessing

```
print(df2.dtypes)
df2["Release date"]=df2["Release date"].astype("datetime64")
print(df2.dtypes)
```

```
Movie Name            object
Release date          object
Duration             float64
Content Rating        object
Production Company    object
Genre                 object
dtype: object
Movie Name                   object
Release date          datetime64[ns]
Duration                    float64
Content Rating               object
Production Company           object
Genre                        object
dtype: object
```

Converted ‘Release date’ to datetime64.

Dataset- 3

```
[49] print("Movies with rotten tomato meter rating of 90% and over:")
     limit=90
     df3=pd.DataFrame({"Movie Name":df[(df['tomatometer_rating']>=limit)].iloc[:,1],"Release date":df[(df['tomatometer_rating']>=limit)].iloc[:,8],"Duration":df[(df['tomatometer_rating']>=limit)].
     print(df3.to_string())
```

```
Movies with rotten tomato meter rating of 90% and over:
                                                                                                Movie Name Release date  Duration Content Rating                          Producti
3                                                                                 12 Angry Men (Twelve Angry Men)   1957-04-13      95.0             NR                         Criterion
6                                                                                                  The 39 Steps   1935-08-01      80.0             NR                 Gaumont British Di
7                                                                                                   3:10 to Yuma   1957-08-07      92.0             NR                          Columbi
11                                                                                                   The Accused   1988-10-14     110.0              R                         Paramour
13                                                                                             The Breaking Point   1950-10-06      97.0             NR                          Warner
14                                                                                                     Adam's Rib   1949-11-18     101.0             NR                      MGM Home Ent
16                                                                                   The Prowler (Cost of Living )   1951-12-03      92.0             PG                          VCI Ent
20                                                                                     The Adventures of Robin Hood   1938-05-14     102.0             PG                              Wa
21                                                                                                       Man Hunt   1941-01-01     100.0             NR                     Twentieth C
31                                                                                                Home of the Brave   2004-05-11      75.0              R                         Emergin
36                                                                                              The Narrow Margin   1952-05-04      71.0             NR                          Warner
46                                                                                                          Alfie   1966-08-24     114.0             PG                         Paramour
50                                                                                                        Aliens   1986-07-18     137.0              R                           20th C
52                                                                                                   All About Eve   1950-01-01     138.0             PG                           20th C
61                                                                                             All the King's Men   1949-11-08     109.0             PG          Sony Pictures Home Ent
62                                                                                                    Street Fight   2005-01-01      86.0             NR               Marshall Curry P
77                                                                                         Four Sheets to the Wind   2007-01-22      81.0              R             First Look Home Ent
88                                                                                                     Borderland   2007-11-09     104.0              R                           After
92                                                                                                 The Cool School   2008-03-07      86.0             NR                            Arth
95                                                                                                      Deep Water   2006-09-03      93.0             PG
112                                                                                                       King Corn   2007-10-12      90.0             NR                         Balcony
135                                                                                                    The Freshman   1990-07-20     102.0             PG          Sony Pictures Home Ent
137                                                                                                           Kenny   2008-03-21     100.0          PG-13                            Xeno
140                                                                   La Graine et le Mulet (The Secret of the Grain) (Couscous)   2007-09-03     151.0             NR
160                                                                                                           Sugar   2008-01-21     118.0              R                             Son
210                                                                                                  Anna Karenina   1935-01-01      95.0             NR                 WARNER BROTHER
216                                                                                             Boogie (Summer Holiday)   2008-05-16     102.0             NR
```

Created another data frame for movies with a tomatometer rating (film critics based) of 90% and above.

Preprocessing

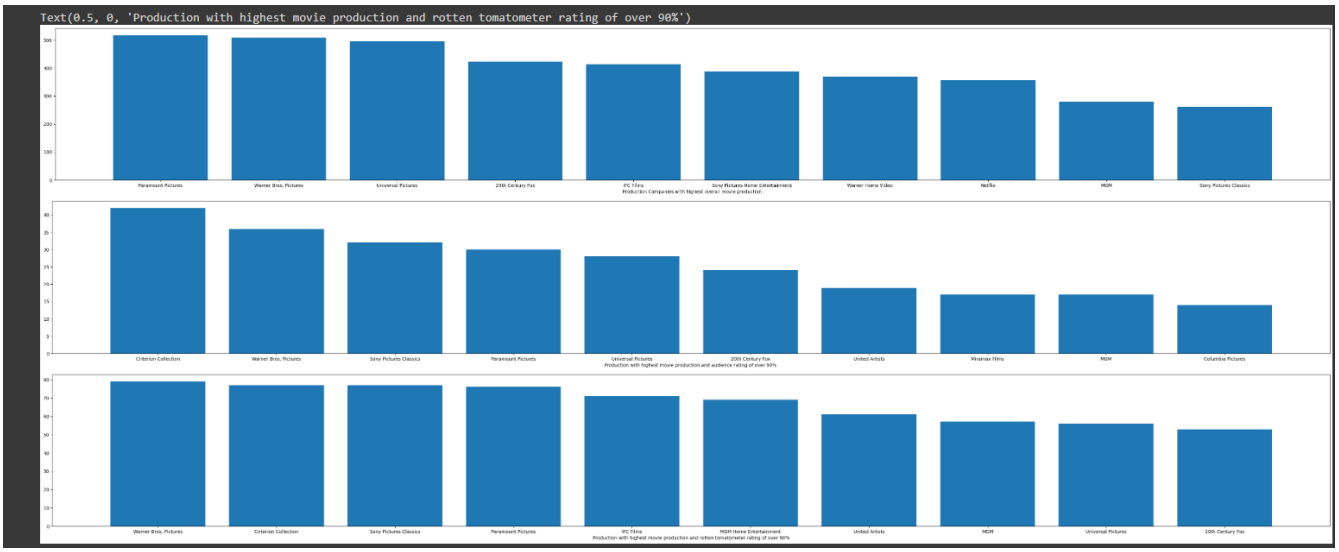
```
print(df3.dtypes)
df3["Release date"]=df3["Release date"].astype("datetime64")
print(df3.dtypes)
```

```
Movie Name             object
Release date           object
Duration              float64
Content Rating         object
Production Company     object
Genre                  object
dtype: object
Movie Name                   object
Release date          datetime64[ns]
Duration                     float64
Content Rating                object
Production Company            object
Genre                         object
dtype: object
```

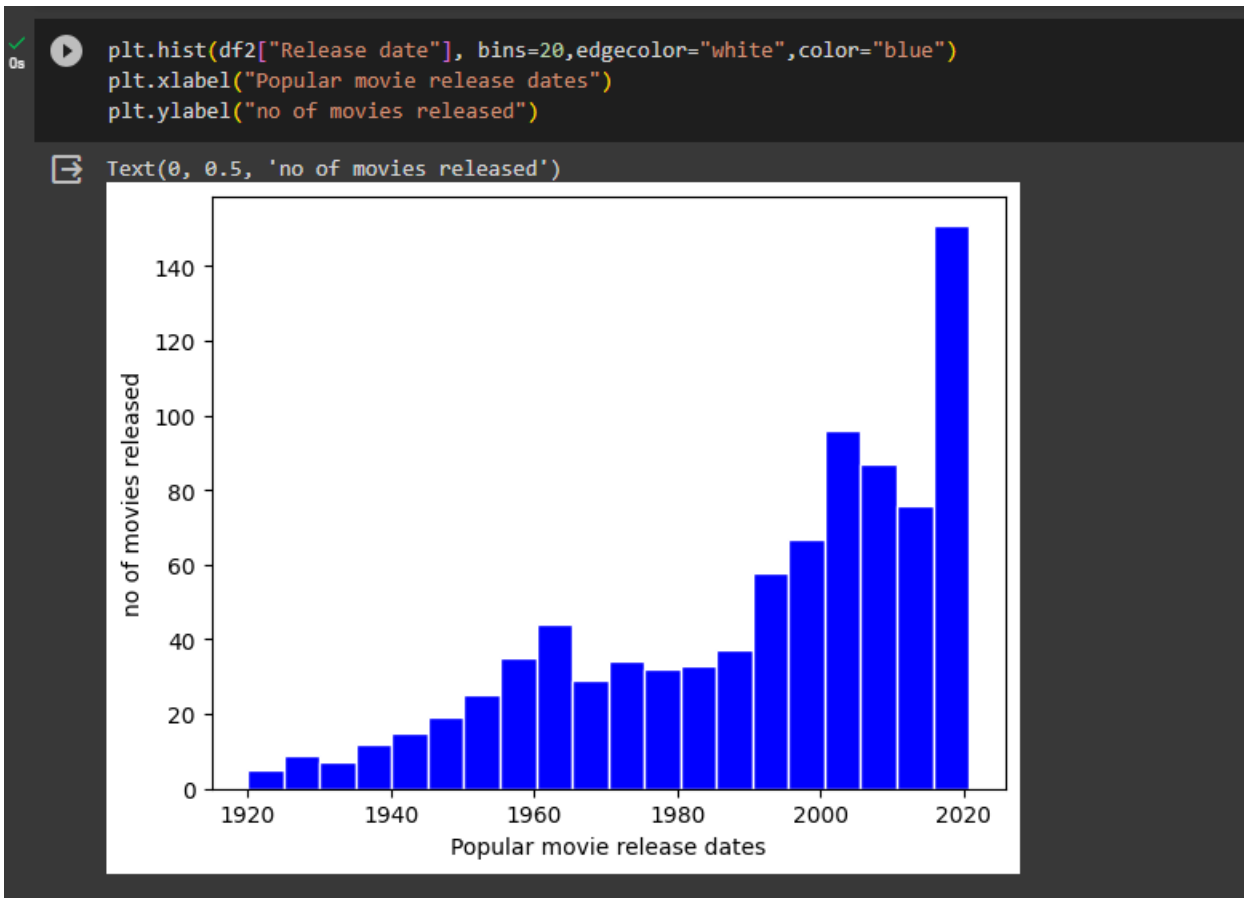
Converted ‘Release date’ to datetime64.

Data Visualization

```
plt.figure(figsize=(50, 6))
value_counts = df['production_company'].value_counts()
plt.bar(value_counts.index[0:10], value_counts.values[0:10])
plt.xlabel("Production Companies with highest overall movie production ")

plt.figure(figsize=(50, 6))
value_counts = df2['Production Company'].value_counts()
plt.bar(value_counts.index[0:10], value_counts.values[0:10])
plt.xlabel("Production with highest movie production and audience rating of over 90% ")

plt.figure(figsize=(50, 6))
value_counts = df3['Production Company'].value_counts()
plt.bar(value_counts.index[0:10], value_counts.values[0:10])
plt.xlabel("Production with highest movie production and rotten tomatometer rating of over 90%")
```



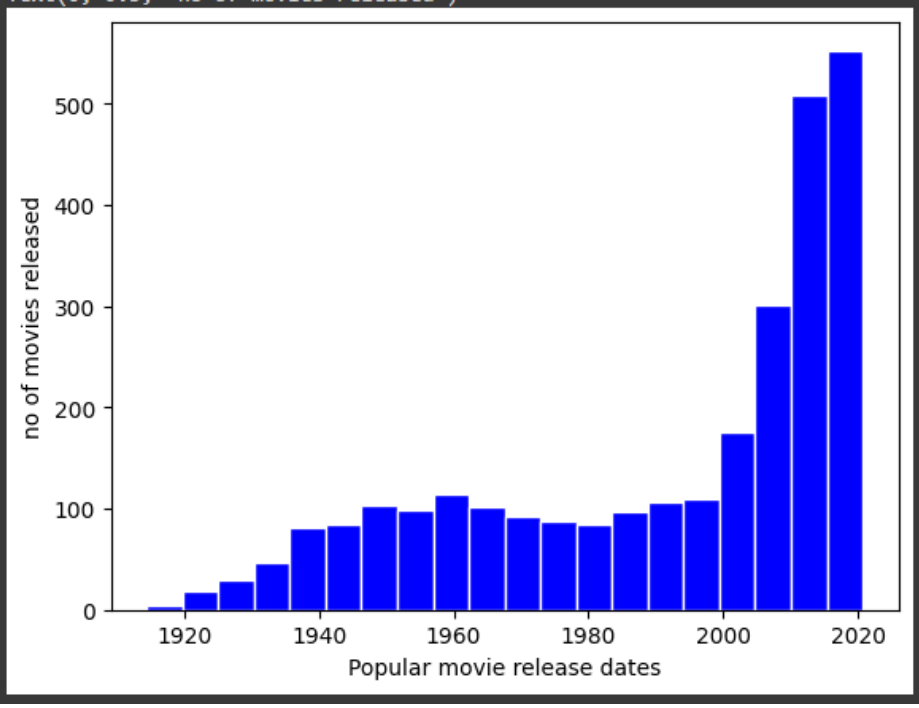
Visually comparing all three data frames which shows the production companies having the number of movies they have produced.



Plotted a histogram to check the number of hit movies that were released. (Audience Rating 90% and above)

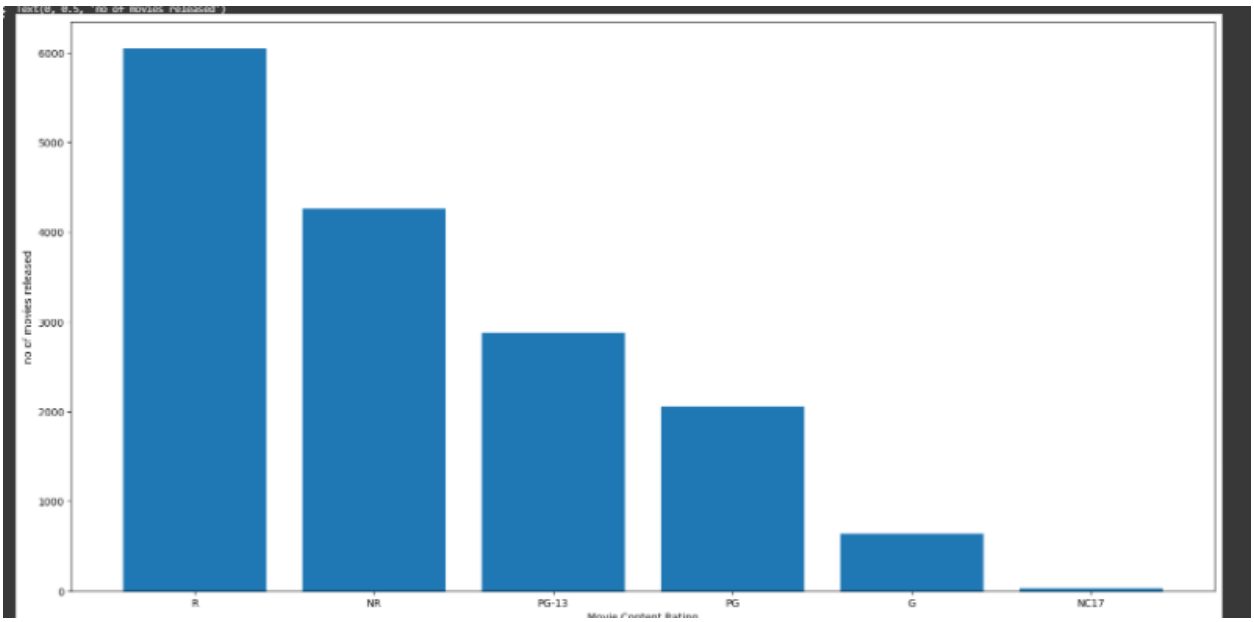
```
plt.hist(df3["Release date"], bins=20,edgecolor="white",color="blue")
plt.xlabel("Popular movie release dates")
plt.ylabel("no of movies released")
```

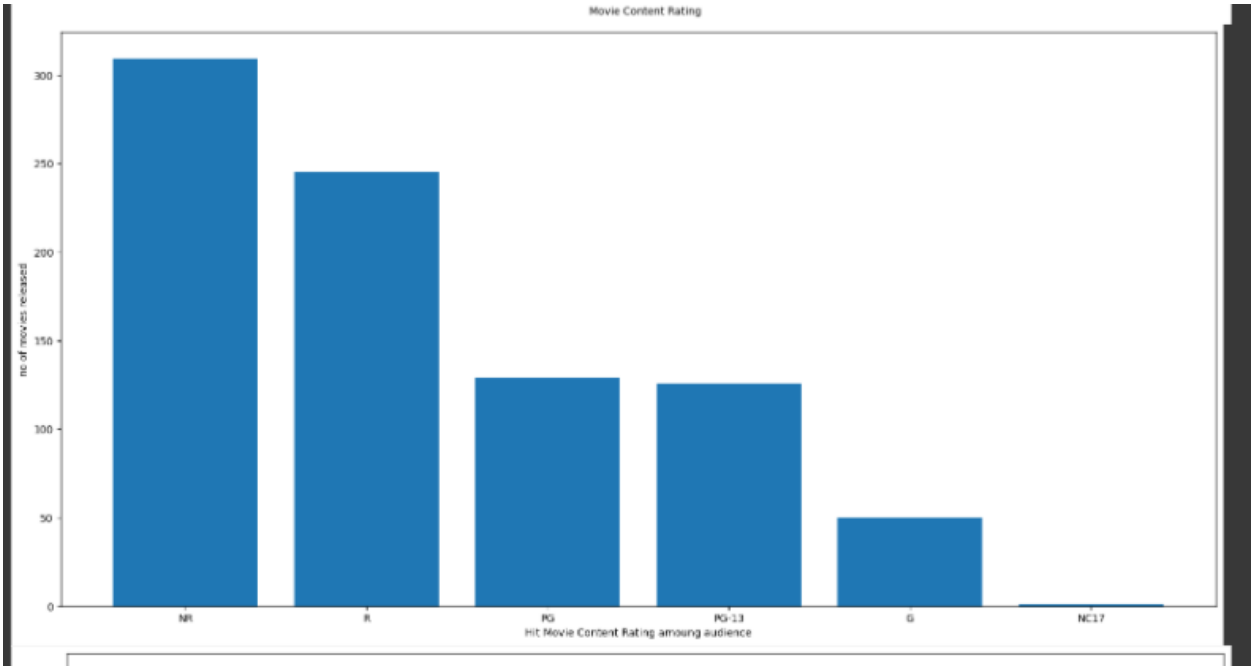
```
Text(0, 0.5, 'no of movies released')
```

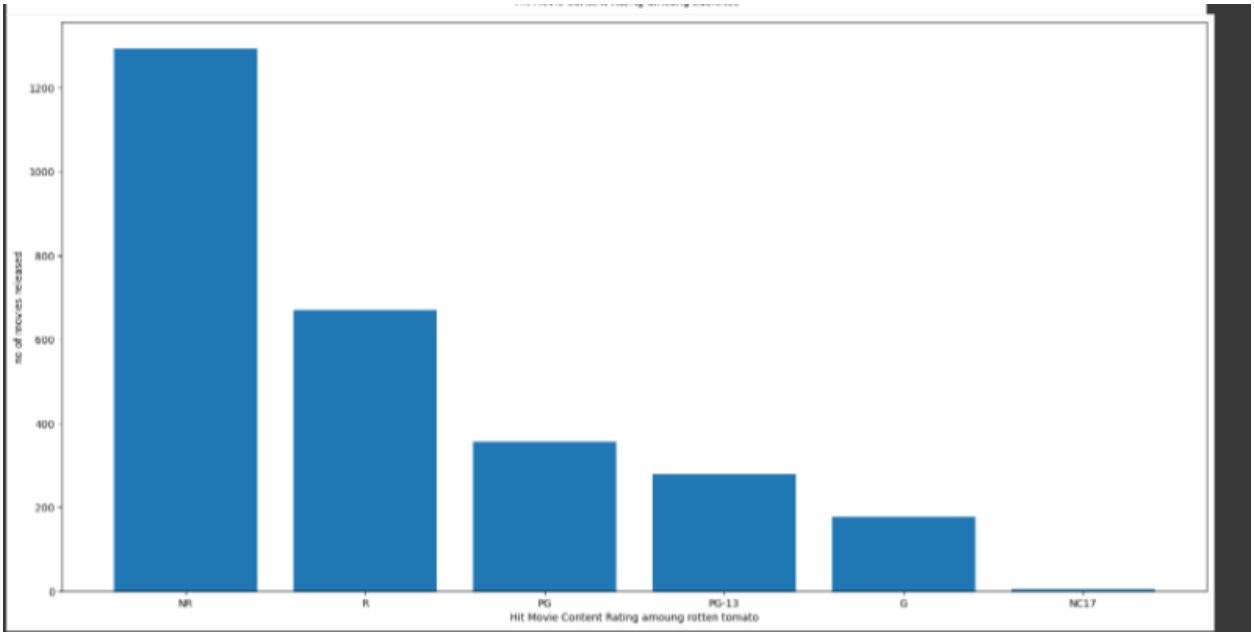


Plotted a histogram to check the number of hit movies that were released. (Tomatometer Rating 90% and above)

```
plt.figure(figsize=(20, 10))
value_counts = df['content_rating'].value_counts()
plt.bar(value_counts.index, value_counts.values)
plt.xlabel("Movie Content Rating")
plt.ylabel("no of movies released")

plt.figure(figsize=(20, 10))
value_counts = df2['Content Rating'].value_counts()
plt.bar(value_counts.index, value_counts.values)
plt.xlabel("Hit Movie Content Rating amoung audience")
plt.ylabel("no of movies released")

plt.figure(figsize=(20, 10))
value_counts = df3['Content Rating'].value_counts()
plt.bar(value_counts.index, value_counts.values)
plt.xlabel("Hit Movie Content Rating amoung rotten tomato")
plt.ylabel("no of movies released")
```

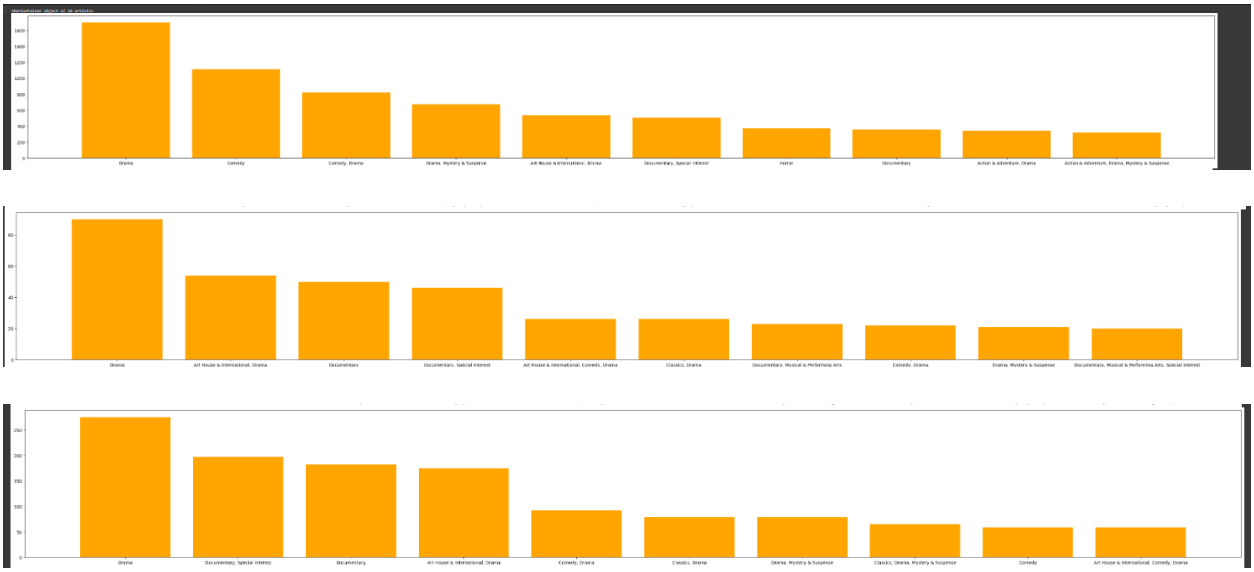

no of movies released
6000
5000
4000
3000
2000
1000
0
R
NR
PG-13
PG
G
NC17


Movie Content Rating
no of movies released
300
250
200
150
100
50
0
NR
R
PG
PG-13
G
NC17
Hit Movie Content Rating amoung audience



Visually comparing all three data frames which shows the content rating of all the movies.

```
plt.figure(figsize=(50, 6))
value_counts = df['genres'].value_counts()
plt.bar(value_counts.index[0:10], value_counts.values[0:10],color="orange")

plt.figure(figsize=(50, 6))
value_counts = df2['Genre'].value_counts()
plt.bar(value_counts.index[0:10], value_counts.values[0:10],color="orange")

plt.figure(figsize=(50, 6))
value_counts = df3['Genre'].value_counts()
plt.bar(value_counts.index[0:10], value_counts.values[0:10],color="orange")
```



Visually comparing all three data frames which shows the genres of all the movies.

Conclusion

We successfully performed exploratory data analysis on a movie dataset about Rotten Tomatoes. First, we explored the dataset and performed various data cleaning techniques. We gained different insights such as identifying the oldest & latest movie, the hit movies, the average movie duration etc. Lastly, we created two data frames (>=90% audience rating & >=90% tomatometer rating) to perform visualization, and compare different aspects of the data.