

```
In [ ]: # Importing the necessary library
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
```

```
In [ ]: df = pd.read_csv(r"C:\Users\foxka\Data Science\Phase 1\Phase 1 Project\title.basics.csv")
df.head(10)
```

Out[]:

	tconst	primary_title	original_title	start_year	runtime_minutes	genres
0	tt0063540	Sunghursh	Sunghursh	2013	175.0	Action,Crime,Drama
1	tt0066787	One Day Before the Rainy Season	Ashad Ka Ek Din	2019	114.0	Biography,Drama
2	tt0069049	The Other Side of the Wind	The Other Side of the Wind	2018	122.0	Drama
3	tt0069204	Sabse Bada Sukh	Sabse Bada Sukh	2018	NaN	Comedy,Drama
4	tt0100275	The Wandering Soap Opera	La Telenovela Errante	2017	80.0	Comedy,Drama,Fantasy
5	tt0111414	A Thin Life	A Thin Life	2018	75.0	Comedy
6	tt0112502	Bigfoot	Bigfoot	2017	NaN	Horror,Thriller
7	tt0137204	Joe Finds Grace	Joe Finds Grace	2017	83.0	Adventure,Animation,Comedy
8	tt0139613	O Silêncio	O Silêncio	2012	NaN	Documentary,History
9	tt0144449	Nema aviona za Zagreb	Nema aviona za Zagreb	2012	82.0	Biography

```
In [ ]: df.tail(10)
```

Out[]:

	tconst	primary_title	original_title	start_year	runtime_minutes	genres
146134	tt9916160	Drømmeland	Drømmeland	2019	72.0	Documentary
146135	tt9916170	The Rehearsal	O Ensaio	2019	51.0	Drama
146136	tt9916186	Illenau - die Geschichte einer ehemaligen Heil...	Illenau - die Geschichte einer ehemaligen Heil...	2017	84.0	Documentary
146137	tt9916190	Safeguard	Safeguard	2019	90.0	Drama,Thriller
146138	tt9916428	The Secret of China	The Secret of China	2019	NaN	Adventure,History,War
146139	tt9916538	Kuambil Lagi Hatiku	Kuambil Lagi Hatiku	2019	123.0	Drama
146140	tt9916622	Rodolpho Teóphilo - O Legado de um Pioneiro	Rodolpho Teóphilo - O Legado de um Pioneiro	2015	NaN	Documentary
146141	tt9916706	Dankyavar Danka	Dankyavar Danka	2013	NaN	Comedy
146142	tt9916730	6 Gunn	6 Gunn	2017	116.0	NaN
146143	tt9916754	Chico Albuquerque - Revelações	Chico Albuquerque - Revelações	2013	NaN	Documentary

In []: `df.duplicated()`

Out[]:

```
0      False
1      False
2      False
3      False
4      False
...
146139  False
146140  False
146141  False
146142  False
146143  False
Length: 146144, dtype: bool
```

In []: `# To find the number of duplicated entries in the data set.`
`df.duplicated().sum()`

Out[]:

```
0
```

In []: `# To find if there are any missing values and the total number.`
`df.isna()`

```
Out[ ]:      tconst primary_title original_title start_year runtime_minutes genres
0      False      False      False      False      False      False
1      False      False      False      False      False      False
2      False      False      False      False      False      False
3      False      False      False      False      False     True      False
4      False      False      False      False      False      False
...
146139  False      False      False      False      False      False
146140  False      False      False      False     True      False
146141  False      False      False      False     True      False
146142  False      False      False      False      False     True
146143  False      False      False      False      True      False
```

146144 rows × 6 columns

```
In [ ]: df.isna().sum()
```

```
Out[ ]: tconst          0
primary_title      0
original_title     21
start_year         0
runtime_minutes   31739
genres            5408
dtype: int64
```

```
In [ ]: df.duplicated().value_counts()
```

```
Out[ ]: False    146144
dtype: int64
```

```
In [ ]: df.tail(2)
```

Out[]:	tconst	primary_title	original_title	start_year	runtime_minutes	genres
146142	tt9916730	6 Gunn	6 Gunn	2017	116.0	NaN
146143	tt9916754	Chico Albuquerque - Revelações	Chico Albuquerque - Revelações	2013	NaN	Documentary

```
In [ ]: df["primary_title"].str.strip()
```

```
Out[ ]: 0 Sunghursh
1 One Day Before the Rainy Season
2 The Other Side of the Wind
3 Sabse Bada Sukh
4 The Wandering Soap Opera
...
146139 Kuambil Lagi Hatiku
146140 Rodolpho Teóphilo - O Legado de um Pioneiro
146141 Dankyavar Danka
146142 6 Gunn
146143 Chico Albuquerque - Revelações
Name: primary_title, Length: 146144, dtype: object
```

```
In [ ]: # Since 'original_title' column already has missing entries and we already have 'primary_title', let's drop it.
df.drop(columns="original_title")
df.head(2)
```

Out[]:	tconst	primary_title	original_title	start_year	runtime_minutes	genres
0	tt0063540	Sunghursh	Sunghursh	2013	175.0	Action,Crime,Drama
1	tt0066787	One Day Before the Rainy Season	Ashad Ka Ek Din	2019	114.0	Biography,Drama

```
In [ ]: missing_values = df['genres'].isna()
print(missing_values)
```

```
0      False
1      False
2      False
3      False
4      False
...
146139  False
146140  False
146141  False
146142   True
146143  False
Name: genres, Length: 146144, dtype: bool
```

```
In [ ]: rows_with_missing_values = df[missing_values]
print(rows_with_missing_values)
```

```

          tconst           primary_title \
16      tt0187902  How Huang Fei-hong Rescued the Orphan from the...
22      tt0253093                  Gangavataran
35      tt0306058                  Second Coming
40      tt0326592                  The Overnight
44      tt0330811            Regret Not Speaking
...
146088  tt9907396      Footloose in the Cotswolds - Part 1
146089  tt9907608      Footloose in the Cotswolds - Part 2
146107  tt9910922  Doctor Who Augmented Reality: Times Magazine
146129  tt9914942             La vida sense la Sara Amat
146142  tt9916730                      6 Gunn

          original_title  start_year \
16      How Huang Fei-hong Rescued the Orphan from the...        2011
22                  Gangavataran        2018
35                  Second Coming        2012
40                  The Overnight        2010
44            Regret Not Speaking        2011
...
146088      Footloose in the Cotswolds - Part 1        2016
146089      Footloose in the Cotswolds - Part 2        2016
146107  Doctor Who Augmented Reality: Times Magazine        2013
146129             La vida sense la Sara Amat        2019
146142                      6 Gunn        2017

       runtime_minutes  genres
16                 NaN    NaN
22                134.0    NaN
35                 95.0    NaN
40                 88.0    NaN
44                 NaN    NaN
...
146088                118.0    NaN
146089                102.0    NaN
146107                 NaN    NaN
146129                 NaN    NaN
146142                116.0    NaN

```

[5408 rows x 6 columns]

In []: df.head(3)

```
Out[ ]:    tconst      primary_title      original_title  start_year  runtime_minutes      genres
0  tt0063540          Sunghursh        Sunghursh     2013       175.0  Action,Crime,Drama
1  tt0066787  One Day Before the Rainy Season  Ashad Ka Ek Din     2019       114.0  Biography,Drama
2  tt0069049  The Other Side of the Wind  The Other Side of the Wind     2018       122.0      Drama
```

```
In [ ]: df = df.drop(columns="original_title")
df.head(3)
```

```
Out[ ]:    tconst      primary_title  start_year  runtime_minutes      genres
0  tt0063540          Sunghursh     2013       175.0  Action,Crime,Drama
1  tt0066787  One Day Before the Rainy Season     2019       114.0  Biography,Drama
2  tt0069049  The Other Side of the Wind     2018       122.0      Drama
```

```
In [ ]: missing_values = df['genres'].isna()
rows_with_missing_values = df[missing_values]
rows_with_missing_values
```

Out[]:

	tconst	primary_title	start_year	runtime_minutes	genres
16	tt0187902	How Huang Fei-hong Rescued the Orphan from the...	2011	NaN	NaN
22	tt0253093	Gangavataran	2018	134.0	NaN
35	tt0306058	Second Coming	2012	95.0	NaN
40	tt0326592	The Overnight	2010	88.0	NaN
44	tt0330811	Regret Not Speaking	2011	NaN	NaN
...
146088	tt9907396	Footloose in the Cotswolds - Part 1	2016	118.0	NaN
146089	tt9907608	Footloose in the Cotswolds - Part 2	2016	102.0	NaN
146107	tt9910922	Doctor Who Augmented Reality: Times Magazine	2013	NaN	NaN
146129	tt9914942	La vida sense la Sara Amat	2019	NaN	NaN
146142	tt9916730	6 Gunn	2017	116.0	NaN

5408 rows × 5 columns

In []: df.isnull()

```
Out[ ]:
```

```
tconst primary_title start_year runtime_minutes genres
```

	0	False	False	False	False	False
1	1	False	False	False	False	False
2	2	False	False	False	False	False
3	3	False	False	False	True	False
4	4	False	False	False	False	False

146139	146139	False	False	False	False	False
146140	146140	False	False	False	True	False
146141	146141	False	False	False	True	False
146142	146142	False	False	False	False	True
146143	146143	False	False	False	True	False

146144 rows × 5 columns

```
In [ ]:
```

```
# To check which rows has a missing value in the 'genres' column.  
df['genres'].isnull()
```

```
Out[ ]:
```

```
0      False  
1      False  
2      False  
3      False  
4      False  
     ...  
146139    False  
146140    False  
146141    False  
146142     True  
146143    False  
Name: genres, Length: 146144, dtype: bool
```

```
In [ ]:
```

```
print(df[df['genres'].isnull()])
```

```
          tconst           primary_title \
16      tt0187902  How Huang Fei-hong Rescued the Orphan from the...
22      tt0253093           Gangavataran
35      tt0306058        Second Coming
40      tt0326592       The Overnight
44      tt0330811    Regret Not Speaking
...
146088  tt9907396  Footloose in the Cotswolds - Part 1
146089  tt9907608  Footloose in the Cotswolds - Part 2
146107  tt9910922  Doctor Who Augmented Reality: Times Magazine
146129  tt9914942      La vida sense la Sara Amat
146142  tt9916730            6 Gunn

          start_year  runtime_minutes genres
16              2011           NaN   NaN
22              2018         134.0   NaN
35              2012         95.0   NaN
40              2010         88.0   NaN
44              2011           NaN   NaN
...
146088          2016         118.0   NaN
146089          2016         102.0   NaN
146107          2013           NaN   NaN
146129          2019           NaN   NaN
146142          2017         116.0   NaN
```

[5408 rows x 5 columns]

In []: df['genres']

```
Out[ ]: 0      Action,Crime,Drama
1      Biography,Drama
2      Drama
3      Comedy,Drama
4      Comedy,Drama,Fantasy
...
146139          Drama
146140      Documentary
146141          Comedy
146142           NaN
146143      Documentary
Name: genres, Length: 146144, dtype: object
```

In []: # To delete nan values in the 'genre' header.
df.dropna(subset=['genres'], inplace=True)

```
df['genres'].isnull().sum()
```

```
Out[ ]: 0
```

```
In [ ]: df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 140736 entries, 0 to 146143
Data columns (total 5 columns):
 #   Column           Non-Null Count  Dtype  
--- 
 0   tconst            140736 non-null   object 
 1   primary_title     140736 non-null   object 
 2   start_year        140736 non-null   int64  
 3   runtime_minutes   112233 non-null   float64
 4   genres             140736 non-null   object 
dtypes: float64(1), int64(1), object(3)
memory usage: 6.4+ MB
```

```
In [ ]: # Now to remove the column 'runtime_minutes' as I perceive it as unnecessary
df = df.drop(columns='runtime_minutes')
df.head(5)
```

```
Out[ ]:
```

	tconst	primary_title	start_year	genres
0	tt0063540	Sunghursh	2013	Action,Crime,Drama
1	tt0066787	One Day Before the Rainy Season	2019	Biography,Drama
2	tt0069049	The Other Side of the Wind	2018	Drama
3	tt0069204	Sabse Bada Sukh	2018	Comedy,Drama
4	tt0100275	The Wandering Soap Opera	2017	Comedy,Drama,Fantasy

```
In [ ]: df.describe()
```

```
Out[ ]:          start_year
count    140736.000000
mean     2014.613986
std      2.735569
min     2010.000000
25%    2012.000000
50%    2015.000000
75%    2017.000000
max     2115.000000
```

```
In [ ]: df.isnull().sum()
```

```
Out[ ]: tconst      0
primary_title  0
start_year     0
genres        0
dtype: int64
```

```
In [ ]: df.nunique()
```

```
Out[ ]: tconst      140736
primary_title  131336
start_year     19
genres        1085
dtype: int64
```

```
In [ ]: df = df.drop_duplicates()
```

```
In [ ]: df.nunique()
```

```
Out[ ]: tconst      140736
primary_title  131336
start_year     19
genres        1085
dtype: int64
```

```
In [ ]: df.dropna(subset=['tconst'], inplace=True)
df['tconst'].isnull().sum()
```

```
Out[ ]: 0
```

```
In [ ]: df.unique()
```

```
Out[ ]: tconst      140736
primary_title   131336
start_year       19
genres          1085
dtype: int64
```

```
In [ ]: # Next, Let's clean the dataset for tmdb.movies.csv
# Let's start by opening up the file.
df2 = pd.read_csv(r"C:\Users\foxka\Data Science\Phase 1\Phase 1 Project\tmdb.movies.csv")
df2.head(3)
```

```
Out[ ]:   Unnamed: 0  genre_ids    id original_language  original_title  popularity  release_date        title  vote_average  vote_count
0            0  [12, 14, 10751]  12444             en  Harry Potter and
                                              the Deathly
                                              Hallows: Part 1  33.533  2010-11-19  Harry Potter
                                                               and the
                                                               Deathly
                                                               Hallows: Part 1  7.7  10788
1            1  [14, 12, 16, 10751]  10191             en  How to Train Your
                                              Dragon  28.734  2010-03-26  How to Train
                                              Your Dragon  7.7  7610
2            2  [12, 28, 878]  10138             en  Iron Man 2  28.515  2010-05-07  Iron Man 2  6.8  12368
```

```
In [ ]: # Now to drop unnecessary columns.
df2 = df2.drop(columns=['genre_ids', 'id', 'release_date'])
df2.head(4)
```

```
Out[ ]:   Unnamed: 0  original_language  original_title  popularity        title  vote_average  vote_count
0            0            en  Harry Potter and the Deathly
                                              Hallows: Part 1  33.533  Harry Potter and the Deathly
                                                               Hallows: Part 1  7.7  10788
1            1            en  How to Train Your Dragon  28.734  How to Train Your Dragon  7.7  7610
2            2            en  Iron Man 2  28.515  Iron Man 2  6.8  12368
3            3            en  Toy Story  28.005  Toy Story  7.9  10174
```

```
In [ ]: df2.isna().sum()
```

```
Out[ ]: Unnamed: 0      0
original_language  0
original_title    0
popularity        0
title             0
vote_average      0
vote_count        0
dtype: int64
```

```
In [ ]: # Now that there are no missing values, let's drop title column.
df2 = df2.drop(columns='title')
df2.head(4)
```

```
Out[ ]:   Unnamed: 0  original_language  original_title  popularity  vote_average  vote_count
0            0           en     Harry Potter and the Deathly Hallows: Part 1  33.533       7.7      10788
1            1           en          How to Train Your Dragon  28.734       7.7      7610
2            2           en            Iron Man 2  28.515       6.8      12368
3            3           en            Toy Story  28.005       7.9      10174
```

```
In [ ]: # Let's drop the first column which is unnamed.
df2 = df2.drop(df2.columns[0], axis = 1)
df2.head(3)
```

```
Out[ ]:  original_language  original_title  popularity  vote_average  vote_count
0           en     Harry Potter and the Deathly Hallows: Part 1  33.533       7.7      10788
1           en          How to Train Your Dragon  28.734       7.7      7610
2           en            Iron Man 2  28.515       6.8      12368
```

```
In [ ]: # To find which movies are popular, let's find the mean for the 'popularity' and 'vote_average' columns.
row_means = df2[['popularity', 'vote_average']].mean(axis=1)
# Let's then add the mean as a column of its own.
df2['average'] = row_means
df2.head(3)
```

```
Out[ ]:    original_language          original_title  popularity  vote_average  vote_count  average
```

	original_language	original_title	popularity	vote_average	vote_count	average
0	en	Harry Potter and the Deathly Hallows: Part 1	33.533	7.7	10788	20.6165
1	en	How to Train Your Dragon	28.734	7.7	7610	18.2170
2	en	Iron Man 2	28.515	6.8	12368	17.6575

```
In [ ]: # After getting the mean between 'popularity' and 'vote_
df2 = df2.sort_values(by='average', ascending=False)
df2.head(10)
```

```
Out[ ]:    original_language          original_title  popularity  vote_average  vote_count  average
```

	original_language	original_title	popularity	vote_average	vote_count	average
23811	en	Avengers: Infinity War	80.773	8.3	13948	44.5365
11019	en	John Wick	78.123	7.2	10081	42.6615
23812	en	Spider-Man: Into the Spider-Verse	60.534	8.4	4048	34.4670
11020	en	The Hobbit: The Desolation of Smaug	53.783	7.3	8392	30.5415
5179	en	The Avengers	50.289	7.6	19673	28.9445
11021	en	Guardians of the Galaxy	49.606	7.9	17958	28.7530
23813	en	Blade Runner 2049	48.571	7.4	6679	27.9855
20617	en	Blade Runner 2049	48.571	7.4	6679	27.9855
23814	en	Fantastic Beasts: The Crimes of Grindelwald	48.508	6.9	4870	27.7040
23815	en	Ralph Breaks the Internet	48.057	7.2	2626	27.6285

```
In [ ]: # Let's open a third dataset that has column headers that are identical to both the first and second dataset.
df3 = pd.read_csv(r"C:\Users\foxka\Data Science\Phase 1\Phase 1 Project\title.ratings.csv")
df3.head(5)
```

```
Out[ ]:      tconst  averagerating  numvotes
0  tt10356526          8.3         31
1  tt10384606          8.9        559
2  tt1042974           6.4         20
3  tt1043726           4.2      50352
4  tt1060240           6.5         21
```

```
In [ ]: df3.isnull().sum()
```

```
Out[ ]: tconst      0
averagerating    0
numvotes        0
dtype: int64
```

```
In [ ]: df3.duplicated().sum()
```

```
Out[ ]: 0
```

```
In [ ]: df3 = df3.rename(columns={'averagerating': 'vote_average', 'numvotes': 'vote_count'})
df3.head(5)
```

```
Out[ ]:      tconst  vote_average  vote_count
0  tt10356526          8.3         31
1  tt10384606          8.9        559
2  tt1042974           6.4         20
3  tt1043726           4.2      50352
4  tt1060240           6.5         21
```

```
In [ ]: # Since we have relabelled the columns from df3 to match the ones in df2, merging can be done.
df2 = df2.merge(df3, on=['vote_average', 'vote_count'])
```

```
In [ ]: df2.duplicated().sum()
```

```
Out[ ]: 11626
```

```
In [ ]: df2 = df2.drop_duplicates()
```

```
In [ ]: df2.duplicated().sum()
```

```
Out[ ]: 0
```

```
In [ ]: df2.head(5)
```

```
Out[ ]:
```

	original_language	original_title	popularity	vote_average	vote_count	average	tconst
0	en	The Mule	33.830	6.5	1584	20.1650	tt1954843
1	en	On the Basis of Sex	32.624	7.4	225	20.0120	tt4029868
2	en	Johnny English Strikes Again	25.478	6.2	1035	15.8390	tt1428453
3	en	The Darkest Minds	23.967	6.8	1158	15.3835	tt6505968
4	en	Kin	23.805	6.2	289	15.0025	tt2065898

```
In [ ]: df = df.rename(columns={'primary_title': 'original_title'})  
df.head(4)
```

```
Out[ ]:
```

	tconst	original_title	start_year	genres
0	tt0063540	Sunghursh	2013	Action,Crime,Drama
1	tt0066787	One Day Before the Rainy Season	2019	Biography,Drama
2	tt0069049	The Other Side of the Wind	2018	Drama
3	tt0069204	Sabse Bada Sukh	2018	Comedy,Drama

```
In [ ]: df2.head(5)
```

Out[]:

	original_language	original_title	popularity	vote_average	vote_count	average	tconst
0	en	The Mule	33.830	6.5	1584	20.1650	tt1954843
1	en	On the Basis of Sex	32.624	7.4	225	20.0120	tt4029868
2	en	Johnny English Strikes Again	25.478	6.2	1035	15.8390	tt1428453
3	en	The Darkest Minds	23.967	6.8	1158	15.3835	tt6505968
4	en	Kin	23.805	6.2	289	15.0025	tt2065898

In []:

```
# Let's merge the 'genre' column on df with the dataset on df2
df2 = df2.merge(df[['tconst', 'genres']], on=['tconst'], how='left')
df2.head(10)
```

Out[]:

	original_language	original_title	popularity	vote_average	vote_count	average	tconst	genres
0	en	The Mule	33.830	6.5	1584	20.1650	tt1954843	Comedy,Sci-Fi
1	en	On the Basis of Sex	32.624	7.4	225	20.0120	tt4029868	Documentary,Drama,Fantasy
2	en	Johnny English Strikes Again	25.478	6.2	1035	15.8390	tt1428453	Biography,Drama
3	en	The Darkest Minds	23.967	6.8	1158	15.3835	tt6505968	Drama,Horror,Thriller
4	en	Kin	23.805	6.2	289	15.0025	tt2065898	Drama,Romance
5	en	Kin	23.805	6.2	289	15.0025	tt6947528	Comedy,Music
6	en	Death Race: Beyond Anarchy	23.163	5.5	93	14.3315	tt4169580	Drama
7	en	Death Race: Beyond Anarchy	23.163	5.5	93	14.3315	tt2041335	Comedy,Drama
8	en	Death Race: Beyond Anarchy	23.163	5.5	93	14.3315	tt3413690	Action
9	en	Someone Marry Barry	7.171	5.5	93	6.3355	tt4169580	Drama

In []:

```
# Let's move the "tconst" column to the front.
column_tconst = df2['tconst']
df2 = df2.drop(columns=['tconst'])
df2.insert(0, 'tconst', column_tconst)
df2.head(5)
```

	tconst	original_language	original_title	popularity	vote_average	vote_count	average	genres
0	tt1954843	en	The Mule	33.830	6.5	1584	20.1650	Comedy,Sci-Fi
1	tt4029868	en	On the Basis of Sex	32.624	7.4	225	20.0120	Documentary,Drama,Fantasy
2	tt1428453	en	Johnny English Strikes Again	25.478	6.2	1035	15.8390	Biography,Drama
3	tt6505968	en	The Darkest Minds	23.967	6.8	1158	15.3835	Drama,Horror,Thriller
4	tt2065898	en	Kin	23.805	6.2	289	15.0025	Drama,Romance

```
In [ ]: df2 = df2.drop_duplicates()
```

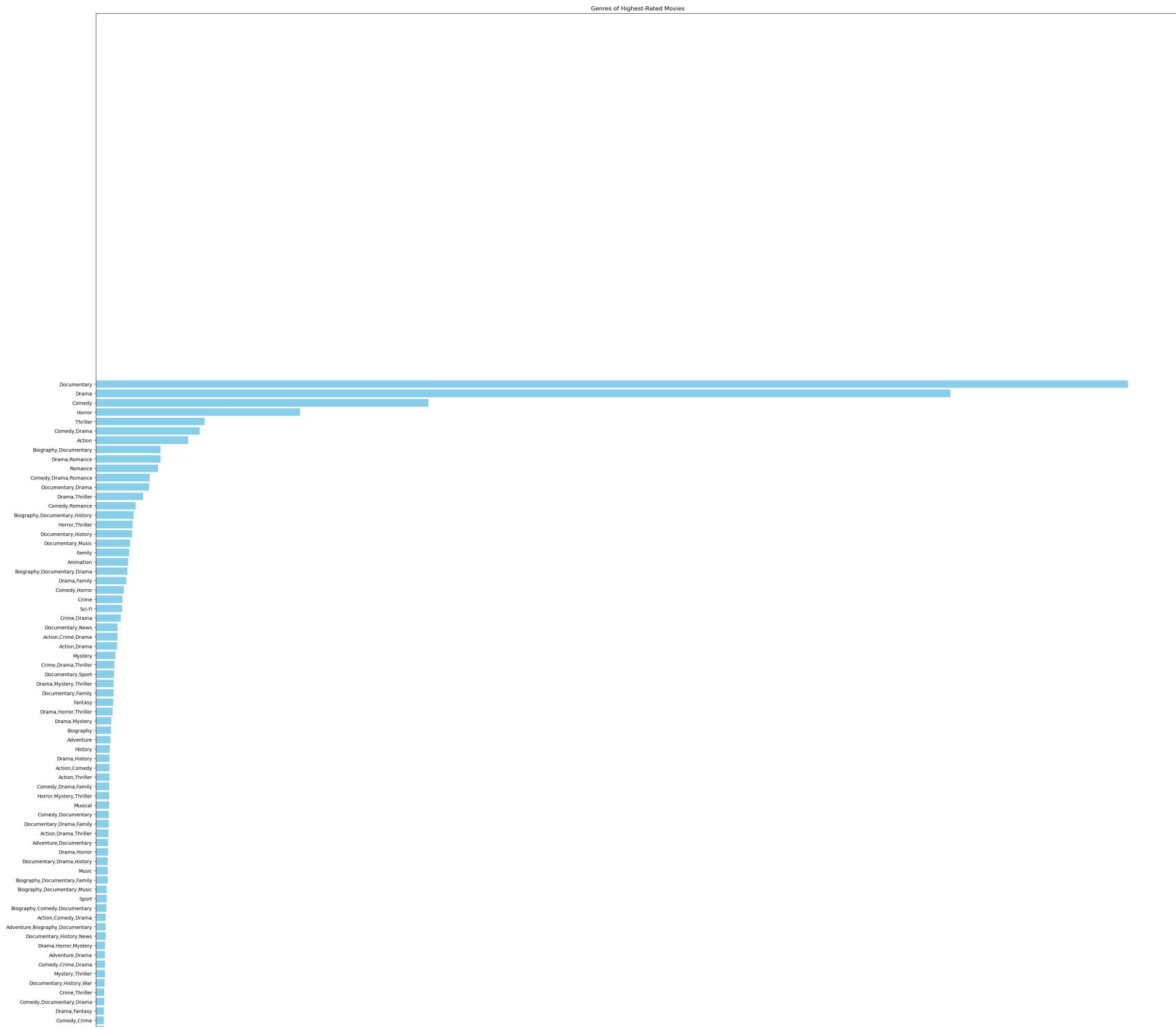
```
In [ ]: df2.duplicated().sum()
```

```
Out[ ]: 0
```

```
In [ ]: df2.isnull().sum()
```

```
Out[ ]: tconst          0
original_language  0
original_title     0
popularity         0
vote_average       0
vote_count         0
average            0
genres             7060
dtype: int64
```

```
In [ ]: df2 = df2.sort_values(by='average', ascending=False)
# We extract the genres associated with the highest-rated movies
highestRatedGenres = df2['genres'].str.split('|').explode()
# Then we measure the frequency of each genre
genreCounts = highestRatedGenres.value_counts().reset_index()
# Then we sort the genres by frequency
genreCounts = genreCounts.sort_values(by='genres', ascending=False)
plt.figure(figsize=(40, 300))
plt.barh(genreCounts['index'], genreCounts['genres'], color='skyblue')
plt.xlabel('Count')
plt.ylabel('Genre')
plt.title('Genres of Highest-Rated Movies')
plt.gca().invert_yaxis()
plt.show()
```



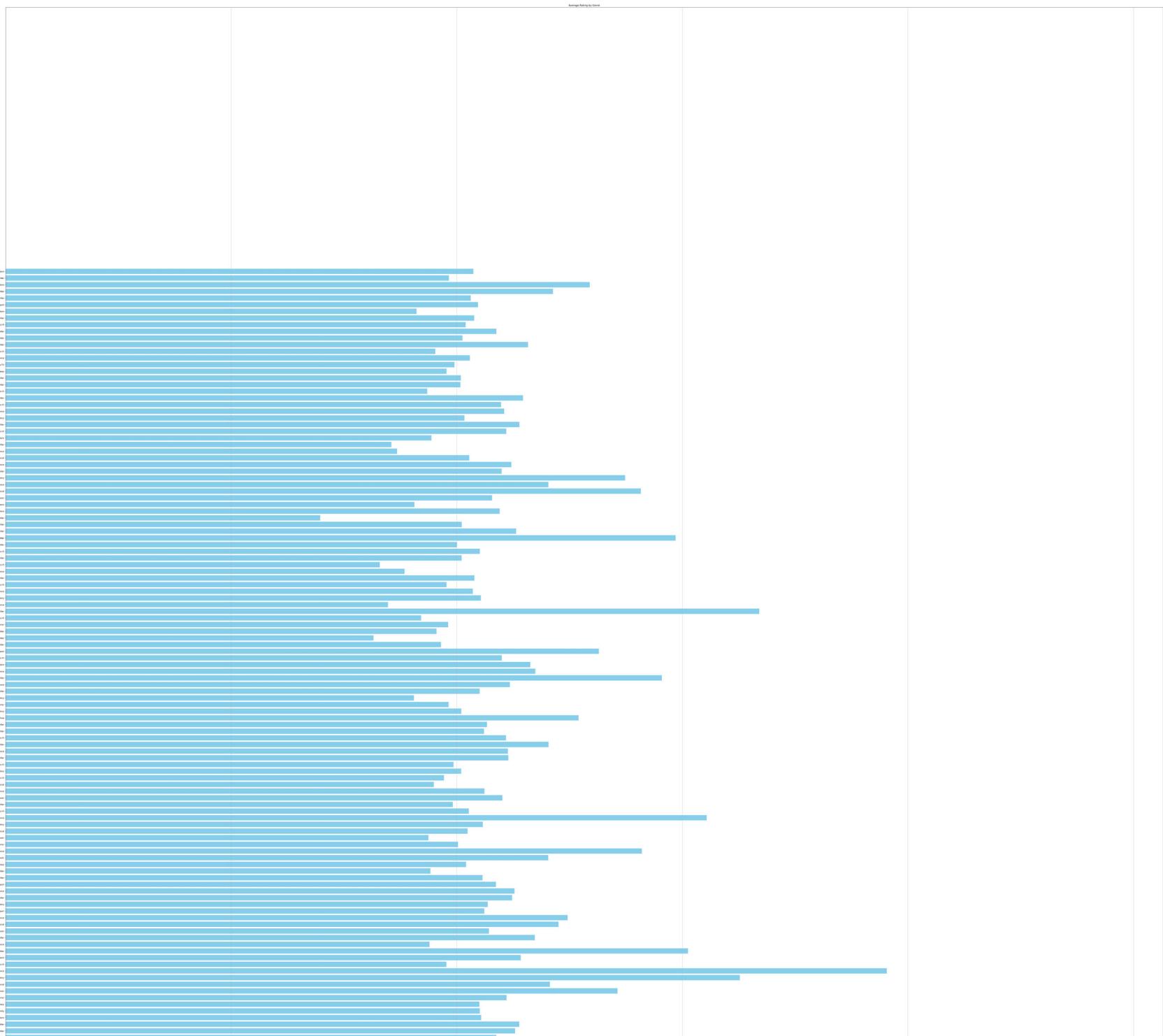
Adventure,Documentary,Drama
Action,Adventure,Comedy
Biography,Drama
Drama,Family,Romance
Drama,Music
Comedy,Family
Documentary,History,Music
Documentary,Drama,News
Action,Comedy,Crime
Documentary,Family,History
Action,Adventure,Drama
Action,Comedy,Humor
Crime,Drama,Mystery
Action,Mystery,Thriller
Drama,Horror,Mystery
Comedy,Drama,Fantasy
Drama,Romance,Thriller
Action,Adventure,Documentary
Documentary,War
Comedy,Horror,Thriller
Drama,Sci-Fi
Crime,Documentary
Comedy,Drama,Humor
Biography,Documentary,Sport
Comedy,Sci-Fi
Adventure,Comedy,Drama
Crime,Documentary,Drama
Action,Crime
Drama,War
Action,Crime,Thriller
Western
Drama,Sport
Adventure,Comedy
Adventure,Documentary,History
Drama,Musical
Action,Humor
Drama,Humor,Romance
Drama,Fantasy,Thriller
War
Biography,Documentary,News
Action,Horror,Science
Comedy,Drama,Music
Adventure,Comedy,Documentary
Comedy,Thriller
Comedy,Drama,Mystery
Crime,Drama,Humor
Drama,Musical,Romance
Comedy,Family,Romance
Comedy,Drama,Thriller
Action,Adventure,Sci-Fi
Crime,Drama,Family
Biography,Crime,Documentary
Action,Adventure,Animation
Horror,Mystery
Action,Comedy,Romance
Action,Adventure,Thriller
Adventure,Family
Action,Romance
Drama,History,War
Action,Horror,Sci-Fi
Fantasy,Humor,Mystery
Horror,Sci-Fi
Adventure,Drama,Fantasy
Crime,Drama,Romance
Comedy,Drama,Musical
Crime,Humor,Thriller
Crime,Mystery,Thriller
Animation,Family
Adventure,Drama,Family
Adventure,Documentary,Thriller
Fantasy,Humor,Thriller
Drama,Fantasy,Sci-Fi
Adventure,Animation,Comedy
Drama,Sci-Fi,Thriller
Adventure,Animation,Fantasy
Romance,Thriller
Sci-Fi,Thriller
Drama,Fantasy,Humor
Comedy,Documentary,History
Animation,Biography,Documentary
Documentary,Drama,Sport
Adventure,Comedy,Family
Comedy,Humor,Sci-Fi
Drama,Music,Romance
Documentary,Humor,Sport
Documentary,Drama,Romance
Documentary,History,Mystery
Comedy,Fantasy,Humor
Documentary,Drama,Music
Comedy,Crime,Thriller
Horror,Sci-Fi,Thriller
Adventure,Documentary,Family
Documentary,Musical
Action,Adventure,Fantasy
Drama,Family,War
Adventure,Documentary,Fantasy
Animation,Comedy,Drama
Action,Comedy,Fantasy
Drama,Family,Mystery
Comedy,Drama,Sci-Fi
Adventure,Comedy,Fantasy
Action,Adventure,Family
Documentary,Humor
Fantasy,Humor
Biography,Drama,History
Action,Sci-Fi,Thriller
Drama,Fantasy,Romance
Comedy,Music
Action,Documentary,GALB

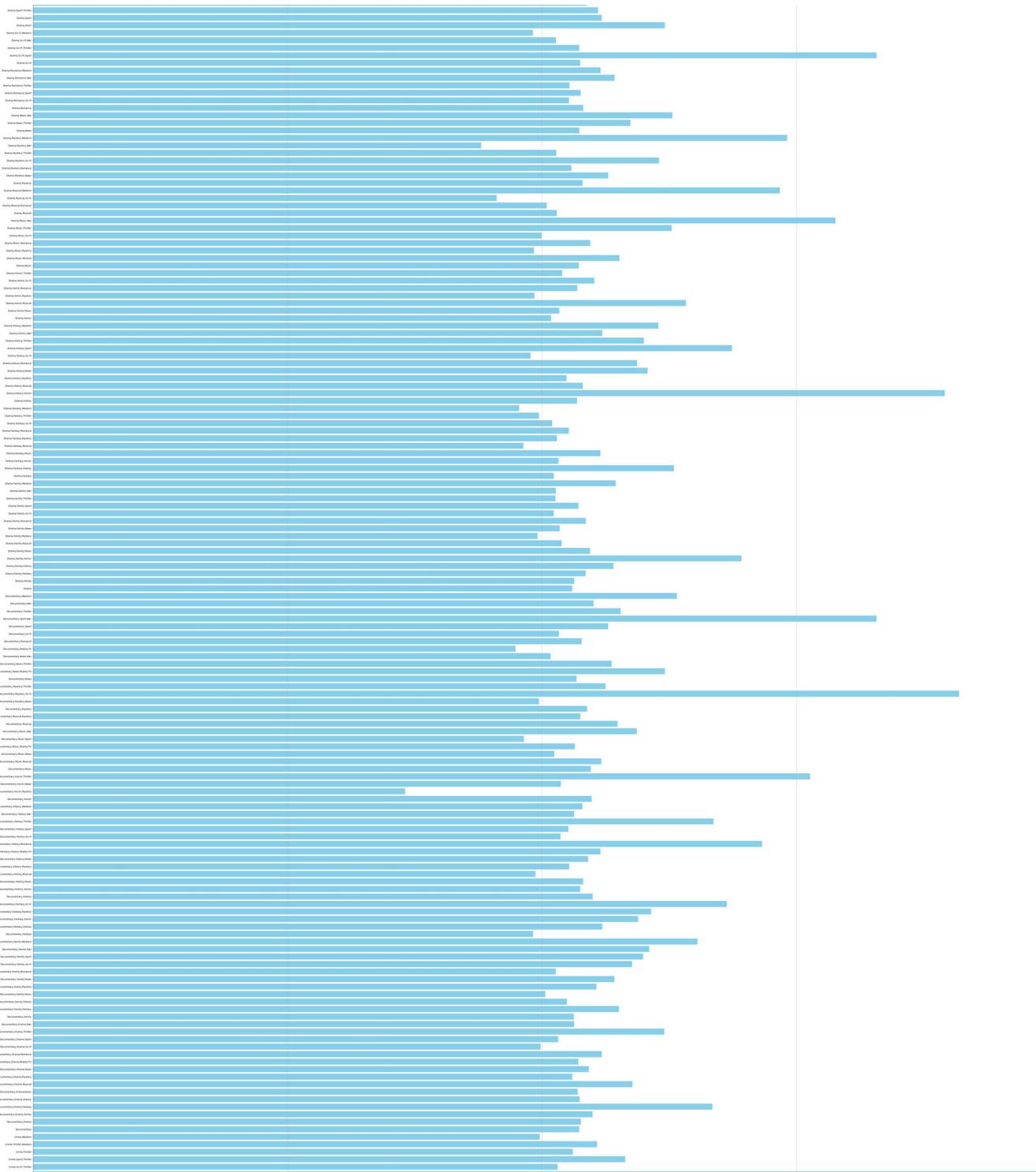
```
Crime,Documentary,History  
Documentary,Drama,War
```

```
In [ ]: # Sort the genres by average rating
genre_ratings = df2.groupby('genres')['average'].mean().reset_index()
# Create a horizontal bar chart with 'genres' on the y-axis
plt.figure(figsize=(100, 500))
plt.barh(genre_ratings['genres'], genre_ratings['average'], color='skyblue')
plt.ylabel('Genre')
plt.xlabel('Average Rating')
plt.title('Average Rating by Genre')
plt.grid(axis='x')

plt.show()
```

```
Action,Adventure,Humor  
Comedy,Drama,History  
Drama,Romance,Sport  
Comedy,Fantasy,Romance  
Animation,Drama  
Action,Adventure  
Action,Drama,History  
Drama,Romance,War  
Action,Fantasy  
Fantasy,Humor,Sci-Fi  
Documentary,Family,News  
Action,Documentary,Drama  
Biography,History  
Mystery,Sci-Fi,Thriller  
Action,Drama,Sport  
Action,Mystery,Sci-Fi  
Romantic,Fantasy  
Documentary,Romance  
Adventure,Family,Fantasy  
Action,Drama,War  
Biography,Drama,Romance  
Action,Drama,Fantasy  
Adventure,Fantasy  
Adventure,Drama,Mystery  
Adventure,Drama,Mystery  
Comedy,Horror,Fantasy  
Crime,Horror,Mystery  
Action,Comedy,Thriller  
Action,Biography,Documentary  
Crime,Drama,Music  
Action,War  
Mystery,Romance,Thriller  
Adventure,Drama,Romance  
Action,Adventure,Crime  
Comedy,Humor,Romance  
Musical,Romance  
Action,Adventure,History  
Documentary,News,War  
Action,Crime,Humor  
Animation,Comedy  
Action,Animation,Drama  
Animation,Biography,Comedy  
Drama,Humor,Romance  
Biography,Drama,Family  
Horror,Music,Sci-Fi  
Comedy,Crime,Mystery  
Fantasy,Mystery  
Comedy,Documentary,Musical  
Adventure,Comedy,Sci-Fi  
Action,Biography,Drama  
Adventure,Crime,Drama  
Action,Thriller,Fantasy  
Romantic,Musical  
Adventure,Drama,Thriller  
Adventure,Fantasy,Humor  
Family,Mystery  
Comedy,Crime,Romance  
Action,Drama,Music  
Action,Drama,Family  
Action,Animation,Sci-Fi  
Mystery,Romance  
Animation,Comedy,Family  
Comedy,Family,Fantasy  
Crime,Mystery  
Fantasy,Romance  
Adventure,Comedy,Humor  
Animation,Fantasy  
Animation,Comedy,Fantasy  
Documentary,Family,Romance  
Action,Adventure,Biography  
Fantasy,Musical,Sci-Fi  
Drama,History,Sci-Fi  
Documentary,Family,Music  
Biography,Comedy,Music  
Comedy,Mystery  
Drama,Mystery,Sci-Fi  
Comedy,Mystery,Sci-Fi  
Crime,Documentary,News  
Adventure,Drama,Sport
```





In []:

