IMDB Movie Data Analysis using Python

Exploring Insights from Top-Rated Movies and Voter Data

Prepared by: Hiba Talha

Capstone 02

Data Analytics | Cohort 5

PROJECT PROPOSAL

We have the data for the 100 top-rated movies from the past decade along with various pieces of information about the movie, its actors, and the voters who have rated these movies online. In this assignment, We will try to find some interesting insights into these movies and their voters, using Python.

Initial steps involve:

- Importing csv file by utilising pandas library and .read_csv() method
- The file will get converted into the Dataframe
- By deploying various methods and attributes, we can retrieve useful information about the dataset for better understanding

Data Overview

- Data Source: IMDb
- Data Description: 100 top-rated movies from the past decade, including movie details (e.g., title, genre, runtime, release year), actor details (e.g., name, gender), and voter details (e.g., age, gender, rating).
- Data Format: CSV file

Tech-Stack Used

Summary for the numeric columns

	title_year	budget	Gross	actor_1_facebook_likes	actor_2_facebook_likes	actor_3_facebook_likes	IMDb_rating	MetaCritic
count	100.000000	1.000000e+02	1.000000e+02	100.000000	99.000000	98.000000	100.000000	95.000000
mean	2012.820000	7.838400e+07	1.468679e+08	13407.270000	7377.303030	3002.153061	7.883000	78.252632
std	1.919491	7.445295e+07	1.454004e+08	10649.037862	13471.568216	6940.301133	0.247433	9.122066
min	2010.000000	3.000000e+06	2.238380e+05	39.000000	12.000000	0.000000	7.500000	62.000000
25%	2011.000000	1.575000e+07	4.199752e+07	1000.000000	580.000000	319.750000	7.700000	72.000000
50%	2013.000000	4.225000e+07	1.070266e+08	13000.000000	1000.000000	626.500000	7.800000	78.000000
75%	2014.000000	1.500000e+08	2.107548e+08	20000.000000	11000.000000	1000.000000	8.100000	83.500000
max	2016.000000	2.600000e+08	9.366622e+08	35000.000000	96000.000000	46000.000000	8.800000	100.000000

8 rows × 53 columns

Data Analysis:

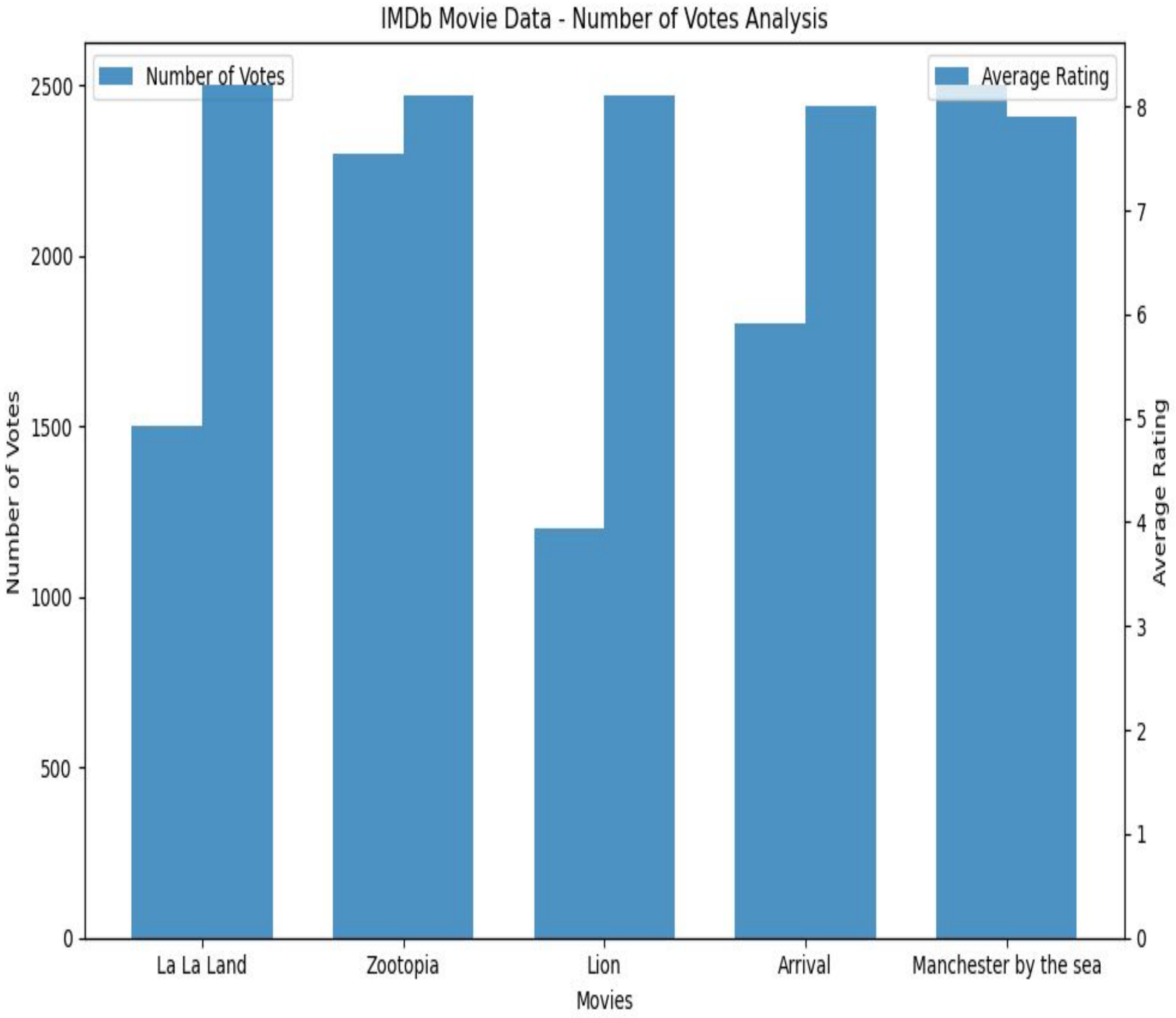
Now let's start with some data manipulation, data analysis, and visualisation to get various insights.

These numbers in the `budget` and `gross` are too big, compromising its readability. Let's convert the unit of the `budget` and `gross` columns from `$` to `million $` first.

	Title	title_year	budget	Gross
0	La La Land	2016	30.0	151.101803
1	Zootopia	2016	150.0	341.268248
2	Lion	2016	12.0	51.738905
3	Arrival	2016	47.0	100.546139
4	Manchester by the Sea	2016	9.0	47.695371

Analysis of IMDb movie data on cleaning the data:

This bar chart shows the analysis of IMDb movie data on the number of votes and average rating for the top 5 movies. The number of votes is plotted on the left y-axis, and the average rating is plotted on the right y-axis. The chart allows us to visually compare the number of votes and average rating for each movie, providing insights into the popularity and quality of these movies based on IMDb user ratings.



Now let's talk about Profit:

Here we find Profit of movies by taking the difference between Gross and budget

	Profit	Gross	budget
0	121.101803	151.101803	30.0
1	191.268248	341.268248	150.0
2	39.738905	51.738905	12.0
3	53.546139	100.546139	47.0
4	38.695371	47.695371	9.0
...	...	...	...
95	9.792000	13.092000	3.3
96	5.114507	8.114507	3.0
97	691.662225	936.662225	245.0
98	146.347721	296.347721	150.0
99	-4.776162	0.223838	5.0

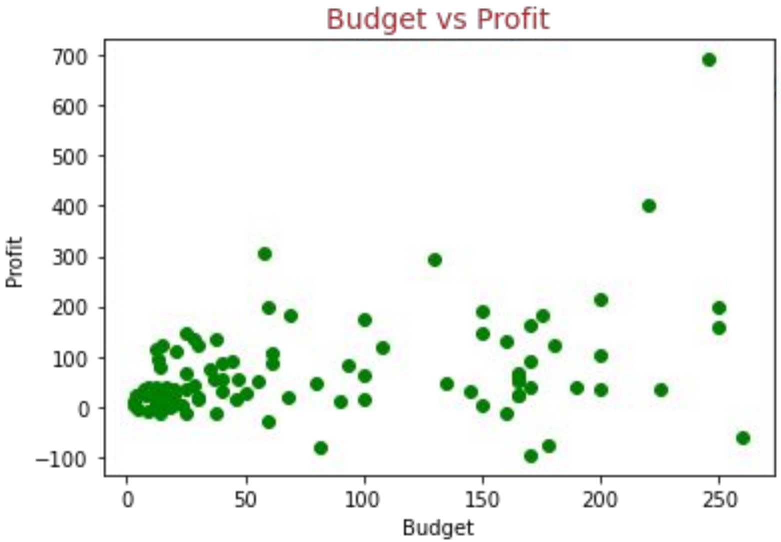
100 rows × 3 columns

Top 5 Profitable movies

	Title	title_year	budget	Gross
97	Star Wars: Episode VII - The Force Awakens	2015	245.0	936.662225
11	The Avengers	2012	220.0	623.279547
47	Deadpool	2016	58.0	363.024263
32	The Hunger Games: Catching Fire	2013	130.0	424.645577
12	Toy Story 3	2010	200.0	414.984497

Budget vs Profit

The dataset contains the 100 best performing movies from the year 2010 to 2016. However scatter plot tells a different story. You can notice that there are some movies with negative profit. Although good movies do incur losses, but there appear to be quite a few movie with losses. What can be the reason behind this? Let's have a closer look at this by finding the movies with negative profit.



Now let's investigate deeply on negative profit movies

	Title	title_year	budget	Gross	a
99	Tucker and Dale vs Evil	2010	5.0	0.223838	
89	Amour	2012	8.9	0.225377	
56	Rush	2013	38.0	26.903709	
66	Warrior	2011	25.0	13.651662	
82	Flipped	2010	14.0	1.752214	

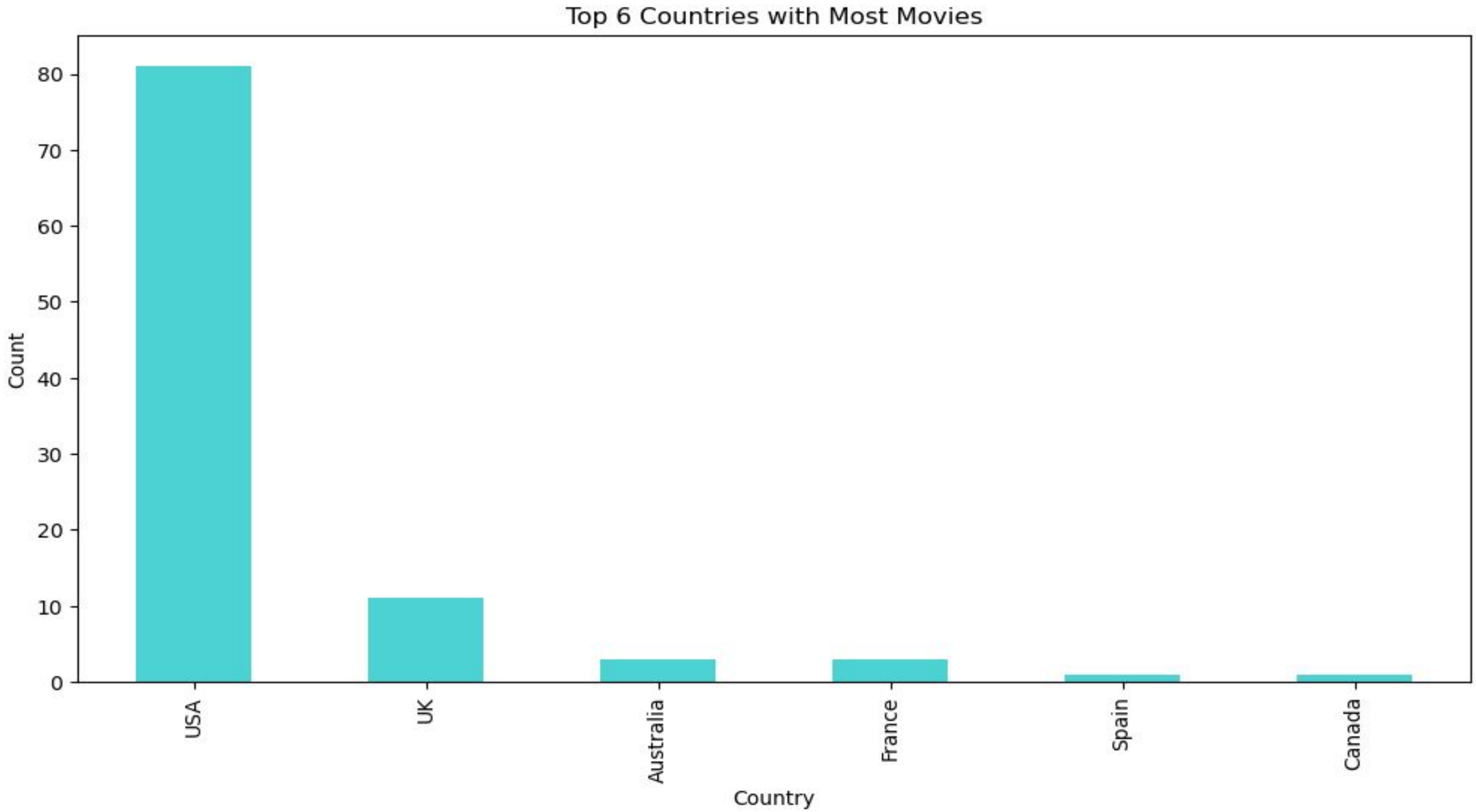
5 rows × 63 columns

The General audience and the critics

You might have noticed the column `MetaCritic` in this dataset. This is a very popular website where an average score is determined through the scores given by the top-rated critics. Second, you also have another column `IMDb_rating` which tells you the IMDb rating of a movie. This rating is determined by taking the average of hundred-thousands of ratings from the general audience.

As a part of this subtask, we are required to find out the highest rated movies which have been liked by critics and audiences alike.

	Title	MetaCritic	IMDb_rating	Avg_rating
95	Whiplash	8.8	8.5	8.65
35	Django Unchained	8.1	8.4	8.25
93	Dallas Buyers Club	8.4	8.0	8.20
97	Star Wars: Episode VII - The Force Awakens	8.1	8.1	8.10
3	Arrival	8.1	8.0	8.05
43	Gone Girl	7.9	8.1	8.00
33	The Martian	8.0	8.0	8.00


Top 6 Countries with Most Movies
Count
0
10
20
30
40
50
60
70
80
USA
UK
Australia
France
Spain
Canada
Country

Find the Most Popular Trios - I

A producer is looking to make a blockbuster movie. There will primarily be three lead roles in his movie and he wish to cast the most popular actors for it. Now, since he don't want to take a risk, he will cast a trio which has already acted in together in a movie before. He want us to find the most popular trio based on the Facebook likes of each of these actors.

The dataframe has three columns to help us out for the same, viz. `actor_1_facebook_likes`, `actor_2_facebook_likes`, and `actor_3_facebook_likes`. Our objective is to find the trios which has the most number of Facebook likes combined. That is, the sum of `actor_1_facebook_likes`, `actor_2_facebook_likes` and `actor_3_facebook_likes` should be maximum.
Lets Find out the top 5 popular trios, and output their names in a list.

We first group all three actor names on the basis of their facebook likes and find out their total likes, and then finally sort them with Top 5 Trios.

actor_1_name	actor_2_name	actor_3_name	actor_1_facebook_likes	actor_2_facebook_likes	actor_3_facebook_likes	Total likes
Dev Patel	Nicole Kidman	Rooney Mara	33000	96000.0	9800.0	138800.0
Leonardo DiCaprio	Tom Hardy	Joseph Gordon-Levitt	29000	27000.0	23000.0	79000.0
Jennifer Lawrence	Peter Dinklage	Hugh Jackman	34000	22000.0	20000.0	76000.0
Casey Affleck	Michelle Williams	Kyle Chandler	518	71000.0	3300.0	74818.0
Tom Hardy	Christian Bale	Joseph Gordon-Levitt	27000	23000.0	23000.0	73000.0

Find the Most Popular Trios - II

In the previous subtask we found the popular trio based on the total number of facebook likes. Let's add a small condition to it ar
that all three actors are popular. The condition is **none of the three actors' Facebook likes should be less than half of the other tv
example, the following is a valid combo:

- actor_1_facebook_likes: 70000
- actor_2_facebook_likes: 40000
- actor_3_facebook_likes: 50000

But the below one is not:

- actor_1_facebook_likes: 70000
- actor_2_facebook_likes: 40000
- actor_3_facebook_likes: 30000

since in this case, `actor_3_facebook_likes` is 30000, which is less than half of `actor_1_facebook_likes`.

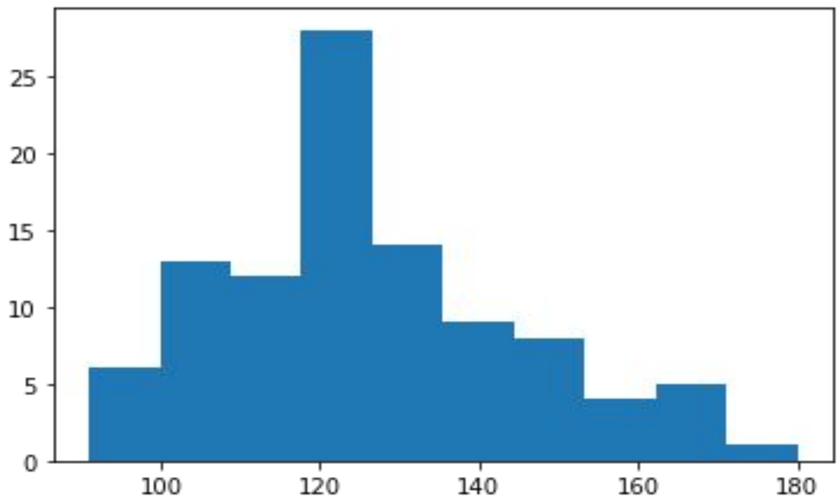
Having this condition ensures that we aren't getting any unpopular actor in our trio (since the total likes calculated in the previous question doesn't tell anything about the individual popularities of each actor in the trio.).

we can do a manual inspection of the top 5 popular trios we have found in the previous subtask and check how many of those trios satisfy this condition. Also, which is the most popular trio after applying the condition above?

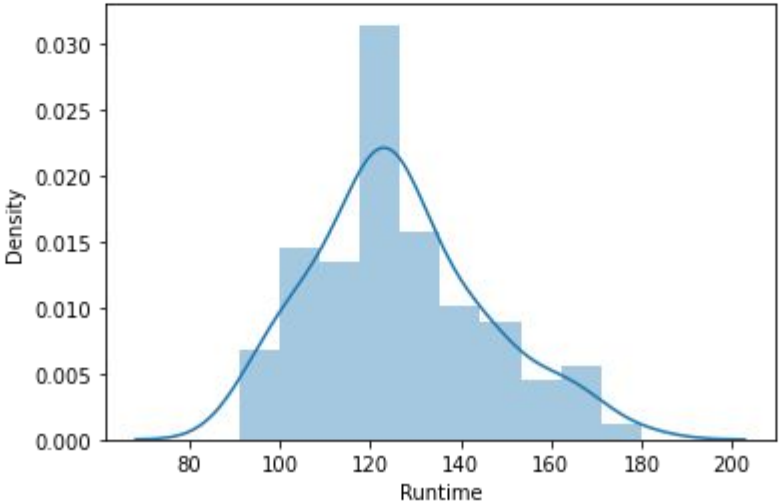
actor_1_name	actor_2_name	actor_3_name	actor_1_facebook_likes	actor_2_facebook_likes	actor_3_facebook_likes	Total likes
Leonardo DiCaprio	Tom Hardy	Joseph Gordon-Levitt	29000	27000.0	23000.0	79000.0
Jennifer Lawrence	Peter Dinklage	Hugh Jackman	34000	22000.0	20000.0	76000.0
Tom Hardy	Christian Bale	Joseph Gordon-Levitt	27000	23000.0	23000.0	73000.0
Chris Hemsworth	Robert Downey Jr.	Scarlett Johansson	26000	21000.0	19000.0	66000.0
Philip Seymour Hoffman	Robin Wright	Brad Pitt	22000	18000.0	11000.0	51000.0

Runtime Analysis

There is a column named `Runtime` in the dataframe which primarily shows the length of the movie. It might be interesting to see how this variable is distributed. Plot a `histogram` or `distplot` of seaborn to find the `Runtime` range most of the movies fall into.

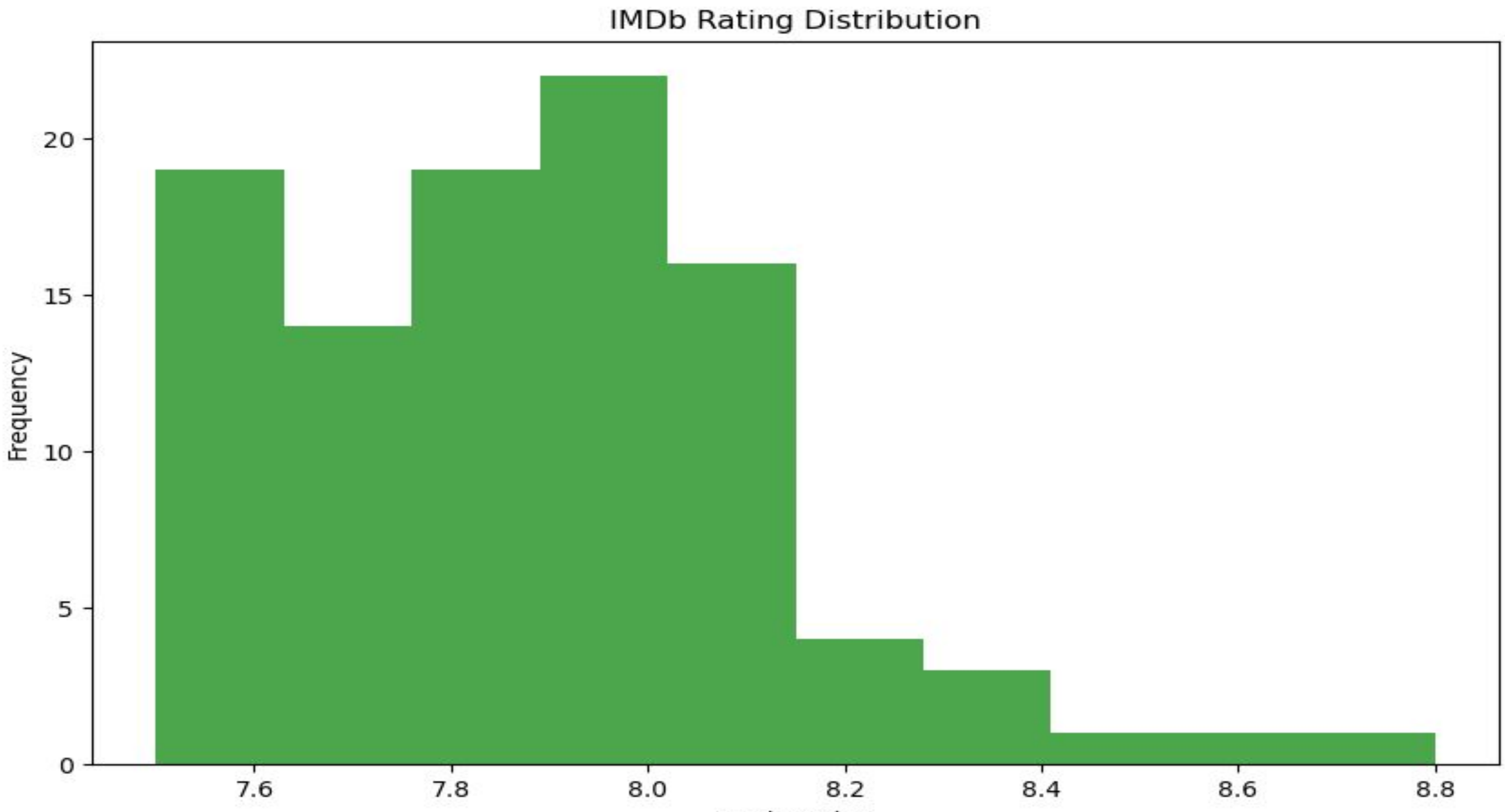


0
5
10
15
20
25
100
120
140
160
180



Density
0.000
0.005
0.010
0.015
0.020
0.025
0.030
80
100
120
140
160
180
200
Runtime

Now Analyze the Data on the basis of IMDb Rating Distribution


IMDb Rating Distribution
Frequency
0
5
10
15
20
7.6
7.8
8.0
8.2
8.4
8.6
8.8

R-Rated Movies

Although R rated movies are restricted movies for the under 18 age group, still there are vote counts from that age group. Among all the R rated movies that have been voted by the under-18 age group, let's find the top 10 movies that have the highest number of votes i.e.`CVotesU18` from the `movies` dataframe. Store these in a dataframe named `PopularR`.

	Title	content_rating	CVotesU18
47	Deadpool	R	4598
36	The Wolf of Wall Street	R	3622
35	Django Unchained	R	3250
29	Mad Max: Fury Road	R	3159
95	Whiplash	R	2878
31	The Revenant	R	2619
40	Shutter Island	R	2321
43	Gone Girl	R	2286
65	The Grand Budapest Hotel	R	2083
72	Birdman or (The Unexpected Virtue of Ignorance)	R	1891

Demographic analysis

If we take a look at the last columns in the dataframe, most of these are related to demographics of the voters. We also have three genre columns indicating the genres of a particular movie. We will extensively use these columns for the third and the final stage of our assignment wherein we will analyse the voters across all demographics and also see how these vary across various genres. So without wasting any time, let's get started with `demographic analysis`.

	genre_1	genre_2	genre_3	MetaCritic	Runtime
97	Action	Adventure	Fantasy	8.1	136
11	Action	Sci-Fi	NaN	6.9	143
47	Action	Adventure	Comedy	6.5	108
32	Action	Adventure	Mystery	7.6	146
12	Animation	Adventure	Comedy	9.2	103

genre_1	MetaCritic	Runtime
Action	192.8	3494
Adventure	86.4	1583
Animation	85.6	1258
Biography	105.2	1666
Comedy	56.9	1064
Crime	70.1	1142
Drama	140.1	2297
Mystery	6.3	138

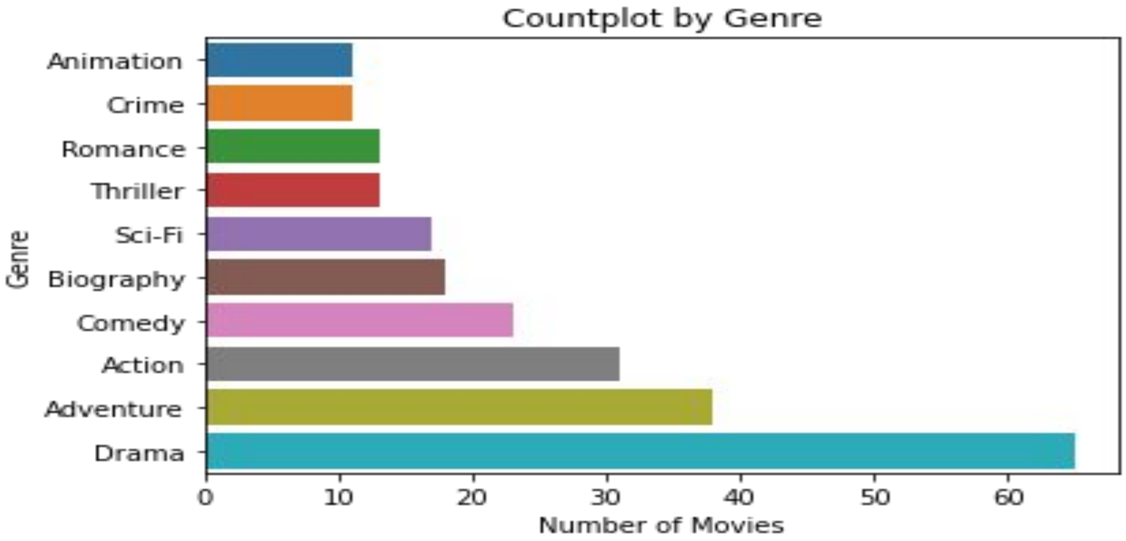
genre_2	MetaCritic	Runtime
Action	30.7	472
Adventure	167.0	2745
Biography	32.2	574
Comedy	54.9	847
Crime	8.2	121
Drama	261.6	4417
Family	6.5	146
Fantasy	14.7	454
History	8.1	142
Horror	6.5	124

genre_3	MetaCritic	Runtime
Adventure	30.7	472
Comedy	54.4	786
Crime	7.5	180
Drama	89.2	1371
Family	8.3	126
Fantasy	21.8	520
History	26.1	409
Music	9.3	128
Mystery	30.0	606
Romance	52.7	963
Sci-Fi	113.8	1968

Here we grouped all three columns of genres. So that the corresponding values of Votes/CVotes get added for each genre.

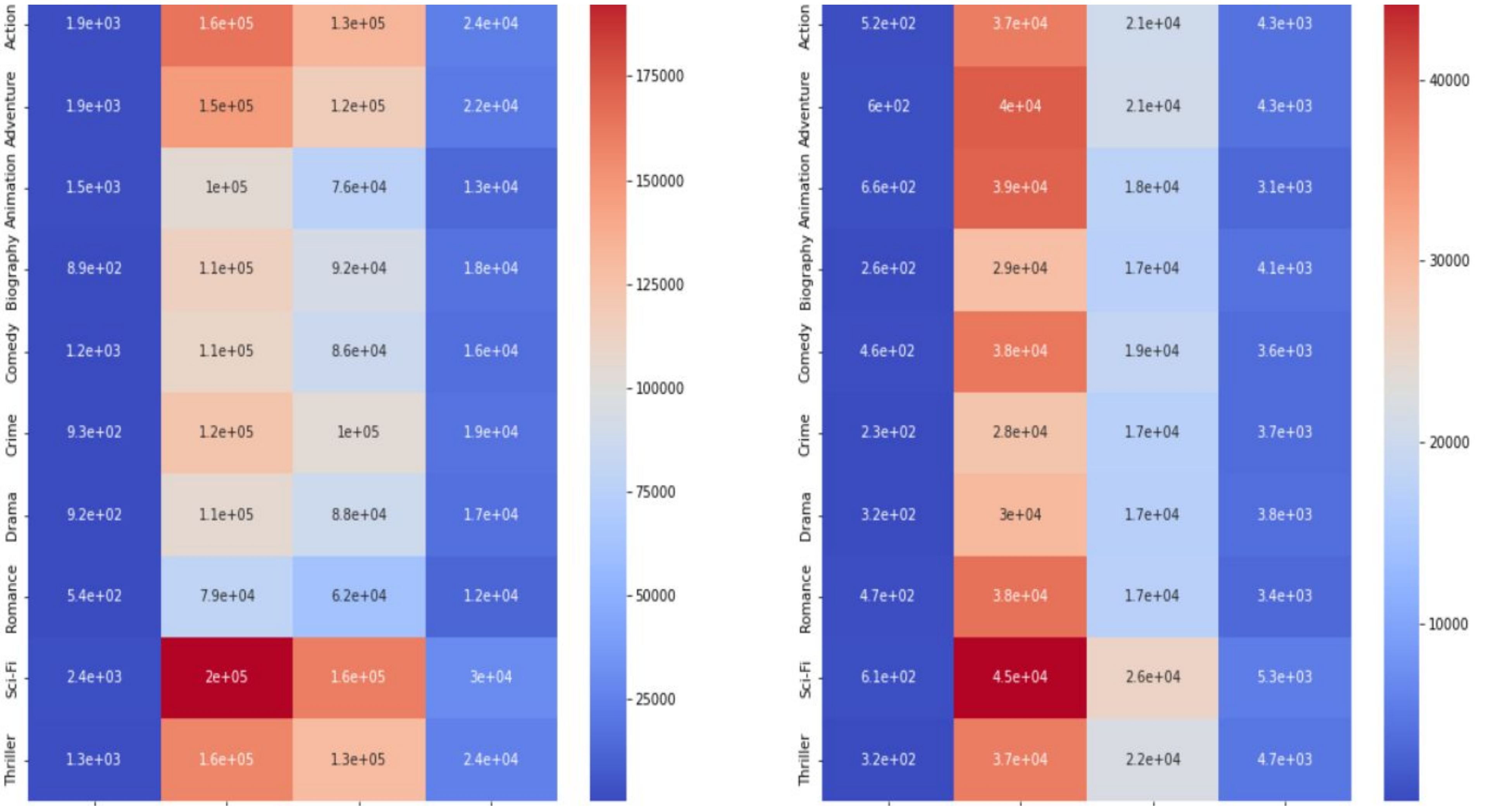
	MetaCritic	Runtime
Action	223.5	3966.0
Adventure	284.1	4800.0
Animation	85.6	1258.0
Biography	137.4	2240.0
Comedy	166.2	2697.0
Crime	85.8	1443.0
Drama	490.9	8085.0
Family	14.8	272.0
Fantasy	36.5	974.0
History	34.2	551.0
Horror	6.5	124.0
Music	18.1	235.0
Musical	6.3	158.0

Countplot by Genre graph



Gender and Genre

Closely looking at the Votes- and CVotes-related columns, We can notice the suffixes `F` and `M` indicating Female and Male. Since we have the vote counts for both males and females, across various age groups, let's see how the popularity of genres vary between the two genders in the dataframe.


Action
Adventure
Animation
Biography
Comedy
Crime
Drama
Romance
Sci-Fi
Thriller
1.9e+03 1.6e+05 1.3e+05 2.4e+04
1.9e+03 1.5e+05 1.2e+05 2.2e+04
1.5e+03 1e+05 7.6e+04 1.3e+04
8.9e+02 1.1e+05 9.2e+04 1.8e+04
1.2e+03 1.1e+05 8.6e+04 1.6e+04
9.3e+02 1.2e+05 1e+05 1.9e+04
9.2e+02 1.1e+05 8.8e+04 1.7e+04
5.4e+02 7.9e+04 6.2e+04 1.2e+04
2.4e+03 2e+05 1.6e+05 3e+04
1.3e+03 1.6e+05 1.3e+05 2.4e+04
175000
150000
125000
100000
75000
50000
25000
Action
Adventure
Animation
Biography
Comedy
Crime
Drama
Romance
Sci-Fi
Thriller
5.2e+02 3.7e+04 2.1e+04 4.3e+03
6e+02 4e+04 2.1e+04 4.3e+03
6.6e+02 3.9e+04 1.8e+04 3.1e+03
2.6e+02 2.9e+04 1.7e+04 4.1e+03
4.6e+02 3.8e+04 1.9e+04 3.6e+03
2.3e+02 2.8e+04 1.7e+04 3.7e+03
3.2e+02 3e+04 1.7e+04 3.8e+03
4.7e+02 3.8e+04 1.7e+04 3.4e+03
6.1e+02 4.5e+04 2.6e+04 5.3e+03
3.2e+02 3.7e+04 2.2e+04 4.7e+03
40000
30000
20000
10000

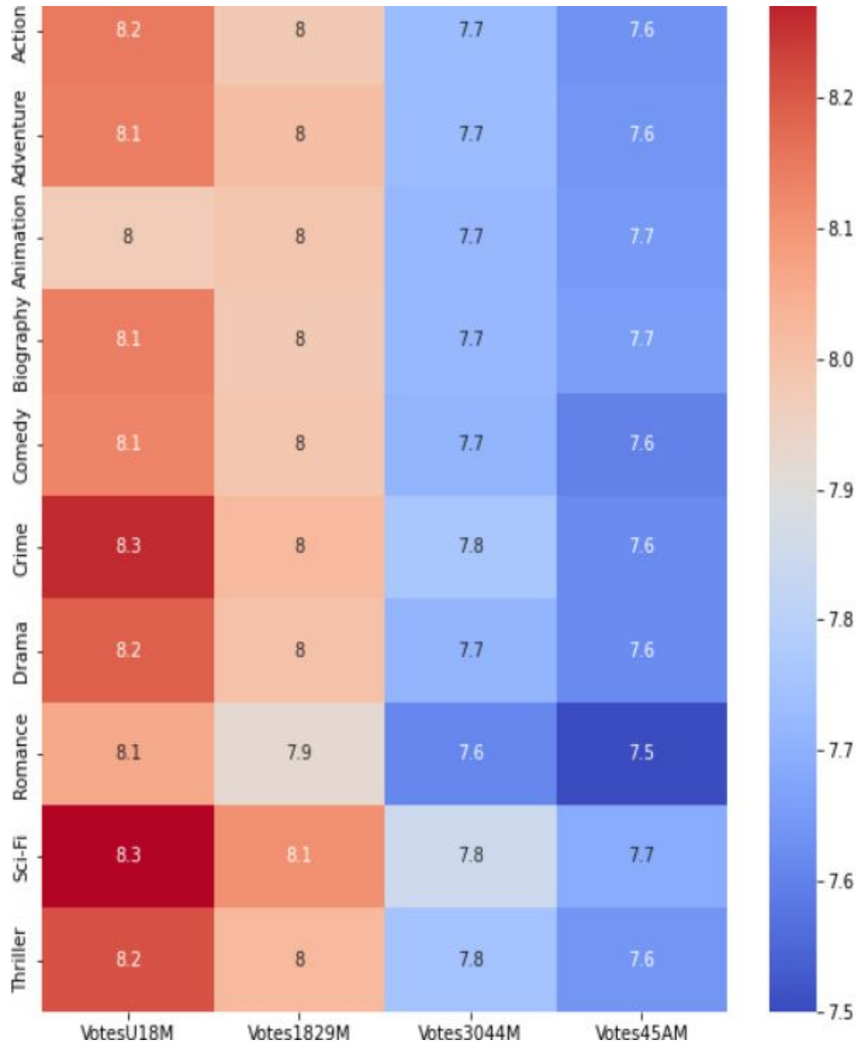

CVotesU18M
CVotes1829M
CVotes3044M
CVotes45AM
CVotesU18F
CVotes1829F
CVotes3044F
CVotes45AF

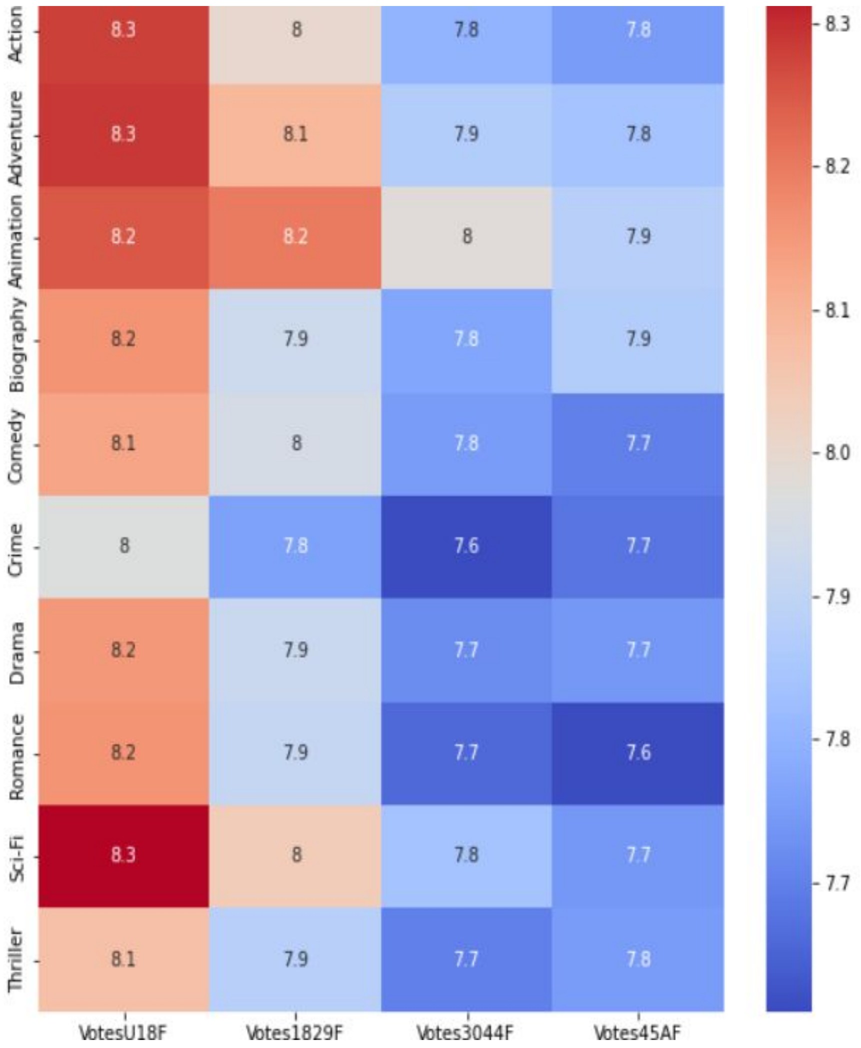
Inferences:

First Heat-Map is CVotes related column

A few inferences that can be seen from the heatmap above is that males have voted more than females, and Sci-Fi appears to be most popular among the 18-29 age group irrespective of their gender. More inferences are:

- Inference 1: Genre romance has got the least number of votes among any age group of males, but there is no such pattern among the females
- Inference 2:Action seems to be the more popular genre among the under 18 males, and Animation appears to be the most popular genre among under 18 females.
- Inference 3: 18-29 age group seems to be most actively voting for any genre irrespective of gender


Action
Adventure
Animation
Biography
Comedy
Crime
Drama
Romance
Sci-Fi
Thriller
VotesU18M
Votes1829M
Votes3044M
Votes45AM
8.2 8 7.7 7.6
8.1 8 7.7 7.6
8 8 7.7 7.7
8.1 8 7.7 7.7
8.1 8 7.7 7.6
8.3 8 7.8 7.6
8.2 8 7.7 7.6
8.1 7.9 7.6 7.5
8.3 8.1 7.8 7.7
8.2 8 7.8 7.6
8.2
8.1
8.0
7.9
7.8
7.7
7.6
7.5


Action
Adventure
Animation
Biography
Comedy
Crime
Drama
Romance
Sci-Fi
Thriller
VotesU18F
Votes1829F
Votes3044F
Votes45AF
8.3 8 7.8 7.8
8.3 8.1 7.9 7.8
8.2 8.2 8 7.9
8.2 7.9 7.8 7.9
8.1 8 7.8 7.7
8 7.8 7.6 7.7
8.2 7.9 7.7 7.7
8.2 7.9 7.7 7.6
8.3 8 7.8 7.7
8.1 7.9 7.7 7.8
8.3
8.2
8.1
8.0
7.9
7.8
7.7

Inferences:

Second Heat-Map is Votes related column

Sci-Fi appears to be the highest rated genre in the age group of U18 for both males and females. Also, females in this age group have rated it a bit higher than the males in the same age group. What more can you infer from the two heatmaps are:

- Inference 1: All genres except for crime and Thriller are more or equally popular amongst U18 female
- Inference 2:18-29 Males likes Scifi more than other genres while females in this group likes animation and adventure more.
- Inference 3: Romance genre is least popular among both genders of 45+ group.

Summary

Firstly, we analyzed the budget vs gross relationship, which revealed that higher budget movies tend to have higher gross earnings, indicating a positive correlation between budget and earnings.

Next, we examined the IMDb rating distribution, which showed that the majority of movies in the dataset have ratings ranging from 7 to 9, with a peak around 8. This suggests that the movies in the top-rated list generally have high IMDb ratings.

In conclusion, the EDA of the IMDb movie data has provided valuable insights into the top-rated movies from the past decade, helping us better understand their budget, IMDb ratings, genre distribution, and country distribution. These findings can be used to inform further analysis and decision-making in the field of movie industry research and provide insights for future movie productions.

THE END

Thank you for your precious time !!!