Predicting Movie Ratings

Matthew Coulombe

Problem Statement

How can we take movie features such as language, violence, drugs, etc. and predict the MPAA rating for that movie?

The Data

The data is sourced from a user on kaggle who scraped content filtering data from VidAngel.

- 3 files
  - movies.csv (1734 unique movies)
  - tags.csv (explanation of the movie tags and categories).
  - movie_tags.csv (23980 unique tags)

Target variable: mpaa_rating from the movies dataset.

Data - movies

	imdb_id	name	title_main	title_subscript	year	mpaa_rating	duration_sec	studio
0	tt11274492	The Out-Laws	The Out-Laws	NaN	2023	R	5700	Happy Madison Productions
1	tt12263384	Extraction 2	Extraction 2	NaN	2023	R	7380	Filmhaus Films|AGBO
2	tt16419074	Air	Air	NaN	2023	R	6720	Mandalay Pictures|Amazon Studios|Skydance Spor...
3	tt14400246	Bird Box Barcelona	Bird Box Barcelona	NaN	2023	TV-MA	7440	Nostromo Pictures|Bluegrass Films|Chris Morgan...
4	tt1745960	Top Gun: Maverick	Top Gun: Maverick	NaN	2022	PG-13	7860	Paramount|Jerry Bruckheimer Films|Don Simpson/...
...	...	...	...	...	...	...	...	...
1729	tt6902676	Guns Akimbo	Guns Akimbo	NaN	2020	R	5700	Ingenious Media|Occupant Films|Four Knights Fi...
1730	tt3813310	Cop Car	Cop Car	NaN	2015	R	5280	Universal
1731	tt2091935	Mr. Right	Mr. Right	NaN	2016	R	5700	Focus World
1732	tt13372794	The Manor	The Manor	NaN	2021	TV-MA	4860	Amazon Studios|Blumhouse Television
1733	tt1464763	Mute	Mute	NaN	2018	R	7560	Netflix

Data - tags

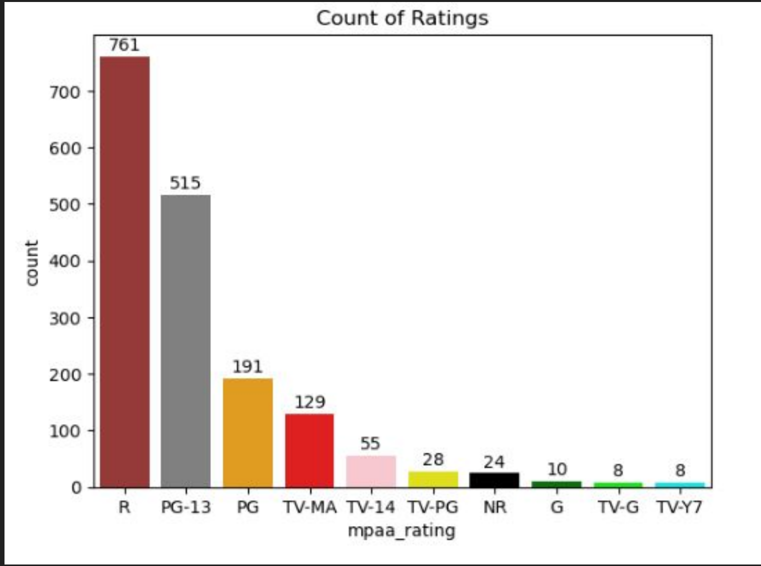
	category	tag_name	title	description
0	language	profanity	Profanity	NaN
1	language	blasphemy	Blasphemy	NaN
2	language	sexual_reference	Sexual References and Innuendos	Any references or jokes about sex, flirting, innuendos, etc.
3	language	childish_language	Childish Language	Generally, things you would not want your 3-year-old to repeat.
4	language	racial_slurs	Racial Slurs and Bigoted Language	Racist, sexist, and/or discriminatory language in any form.
5	violence	non_graphic	Non-Graphic	Violence without blood.
6	violence	graphic	Graphic	Violence with blood or breaking bones.
7	violence	disturbing_images	Disturbing Images	Dead bodies, severed body parts, or object protruding from body
8	violence	gore	Gore	Gore, bloody guts, bloody severed body parts.
9	violence	violence_implied	Implied Violence	The violence is not seen on screen. Graphic descriptions or details of a violent act.
10	immodesty	immodesty	Immodesty	Bikinis, focused chest shots, or bare midriffs for women. Formfitting underwear, breifs, or speedos for men.
11	immodesty	nudity_without_sex	Nudity (without sex)	Skinny dipping, bathing, flashing, mooning, etc.
12	immodesty	nudity_art	Statues and Paintings	Art, statues, mannequins, drawings, stained glass, reliefs, etc.
13	immodesty	nudity_implied	Implied Nudity	Not wearing clothing but private areas are hidden.
14	sexual	sexually_suggestive	Sexually Suggestive	Behaviors with action or sexual undertones enticing or implying sexual intent.
15	sexual	kissing_normal	Normal Kissing	Lip-to-lip kissing.
16	sexual	kissing_passion	Passionate Kissing	French kissing, making out, or sensually kissing parts of the body.
17	sexual	sex_implied	Implied Sex	When sex happens off-screen, or immediately before/after
18	sexual	sexual_assault	Sexual Assault	Rape, attempted rape, bestiality, etc. References to rape or molestation.
19	sexual	sex_without_nudity	Sex without Nudity	Sex with a discreet camera angle, under bedsheets, etc.
20	sexual	sex_with_nudity	Sex with Nudity	Sex shown with any body part that would normally be covered by a bikini or Speedo.
21	drugs	drugs_legal	Legal Use	Legal drinking/smoking
22	drugs	drugs_implied	Implied Use	All discussion, handling, making, and visibility of illegal drugs and underage drinking/smoking
23	drugs	drugs_illegal	Illegal Use	Consumption of illegal drugs and underage drinking/smoking, including in the background.
24	other	bodily_functions	Bodily Functions/Jokes	Gross bodily fluids/functions such as a person passing gas. Nosebleeds and such that are not from violence. Potty talk.
25	other	objectionable	Objectionable/Disturbing/Scary	Violent seizures, condoms, tattoo needles etc.
26	other	vulgar_gestures	Vulgar Gestures	Crotch-grabbing, gestures for profanities, mimicking any sex act, etc.
27	other	medical_graphic	Medical - Graphic	Medical procedures where blood, organs, or anything gross is shown.
28	other	medical_procedures	Medical - Procedures	Vaccines and medical shots when the needle penetrates the skin. Doctor procedures that do not include blood.
29	other	life_events	Life Events	Death by natural/non-violent causes. Birth, labor, and contractions if anything is seen or when they start giving birth.

Data - movie_tags

	imdb_id	category	tag_name	occurrence_cnt	duration_sec
0	tt0052357	language	blasphemy	1	0
1	tt0052357	violence	non_graphic	5	30
2	tt0052357	violence	disturbing_images	1	0
3	tt0052357	immodesty	immodesty	1	6
4	tt0052357	immodesty	nudity_implied	1	30
...	...	...	...	...	...
23975	tt9902160	violence	non_graphic	9	18
23976	tt9902160	violence	graphic	4	12
23977	tt9902160	immodesty	immodesty	3	30
23978	tt9902160	sexual	sexually_suggestive	1	6
23979	tt9902160	other	medical_graphic	1	6

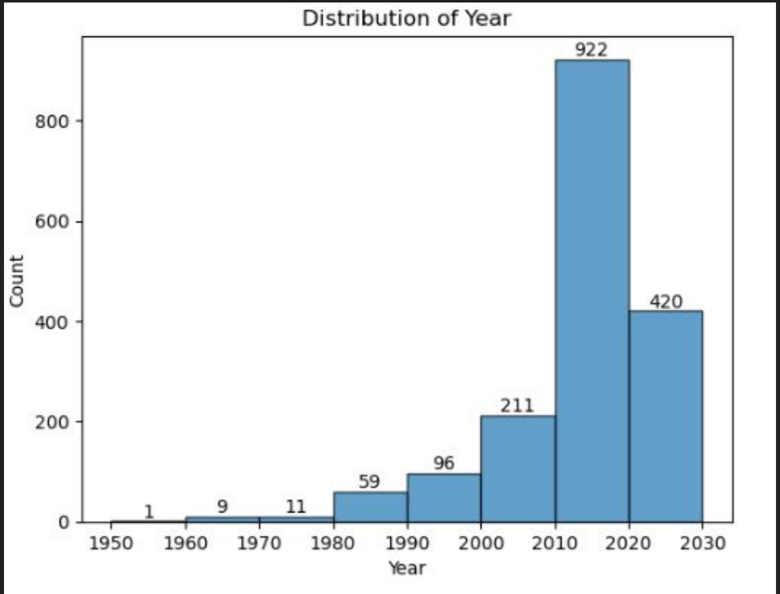
Exploratory Data Analysis [EDA] - Rating (Movies)

1. G
2. PG
3. PG-13
4. R
5. NR
6. TV-G
7. TV-Y7
8. TV-PG
9. TV-14
10. TV-MA



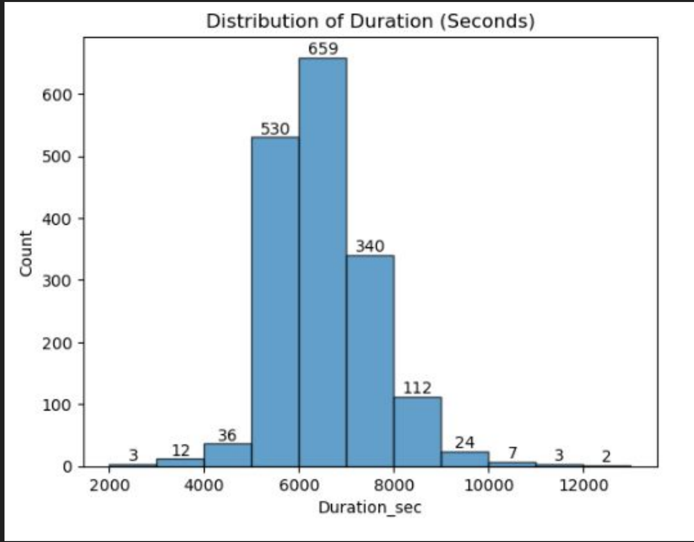
EDA - Year (Movies)

- Year the movie was made
- Skewed toward recent movies

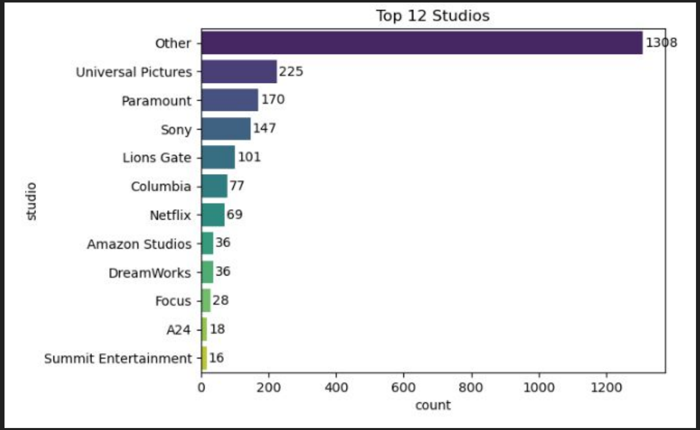


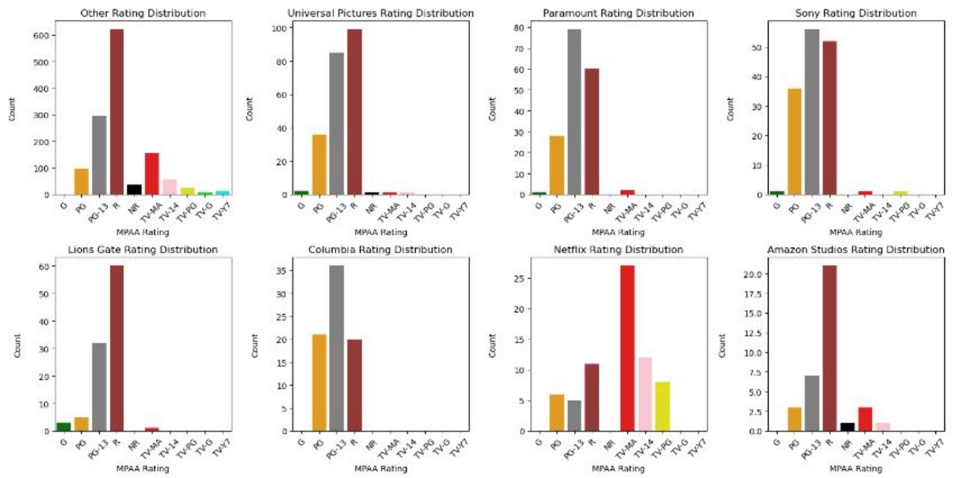
EDA - Duration (Movies)

- Average time around 6500 seconds
- Approximately equal to 1 hour and 48 minutes

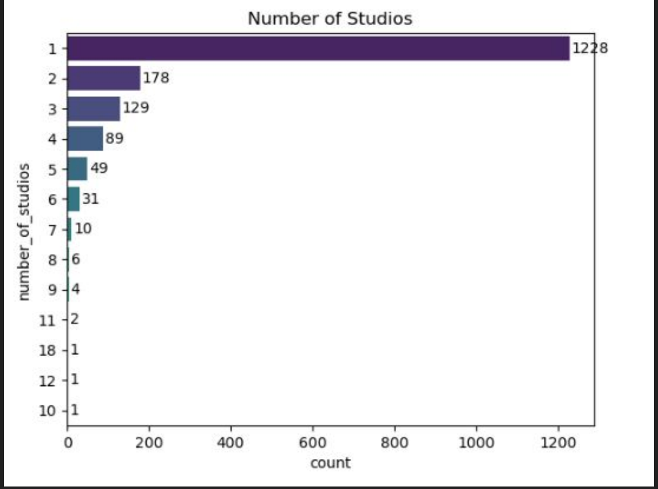


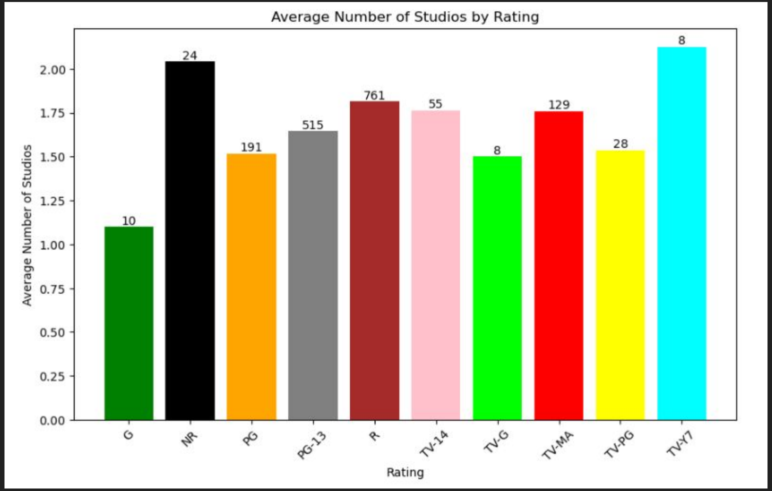
EDA - Studio (Movies)





EDA - Number of Studios (Movies)

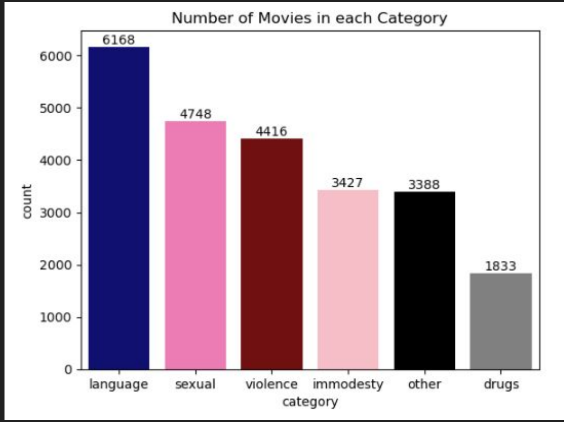




EDA - Category (Movie Tags)

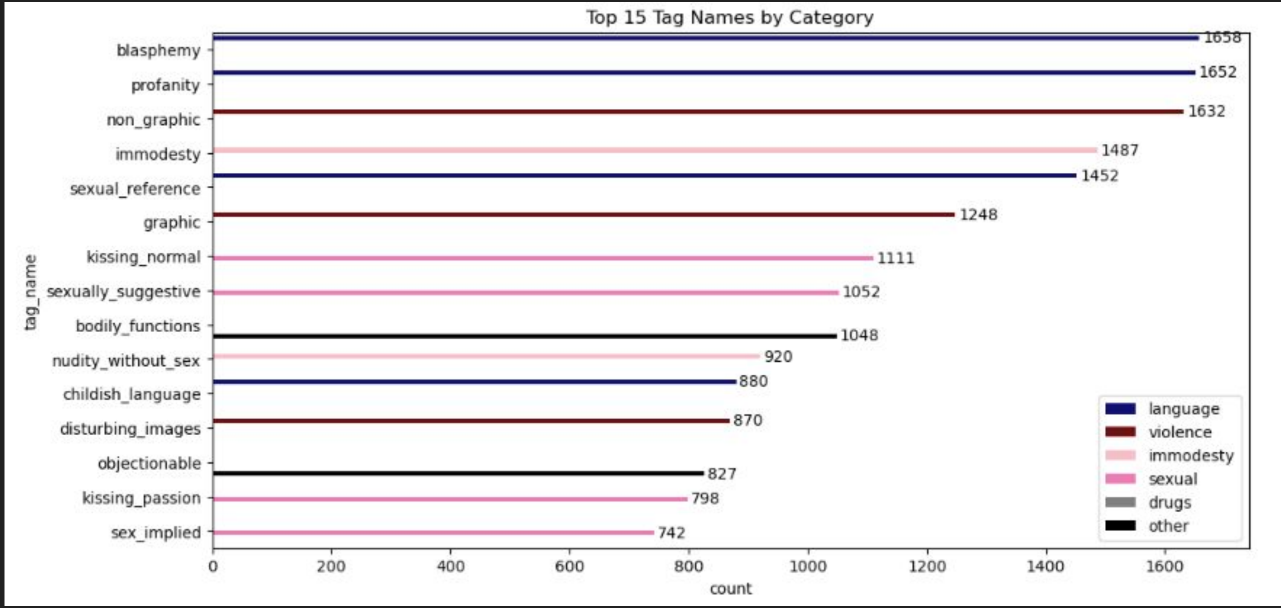
6 Categories

- Language
- Sexual
- Violence
- Immodesty
- Other
- Drugs



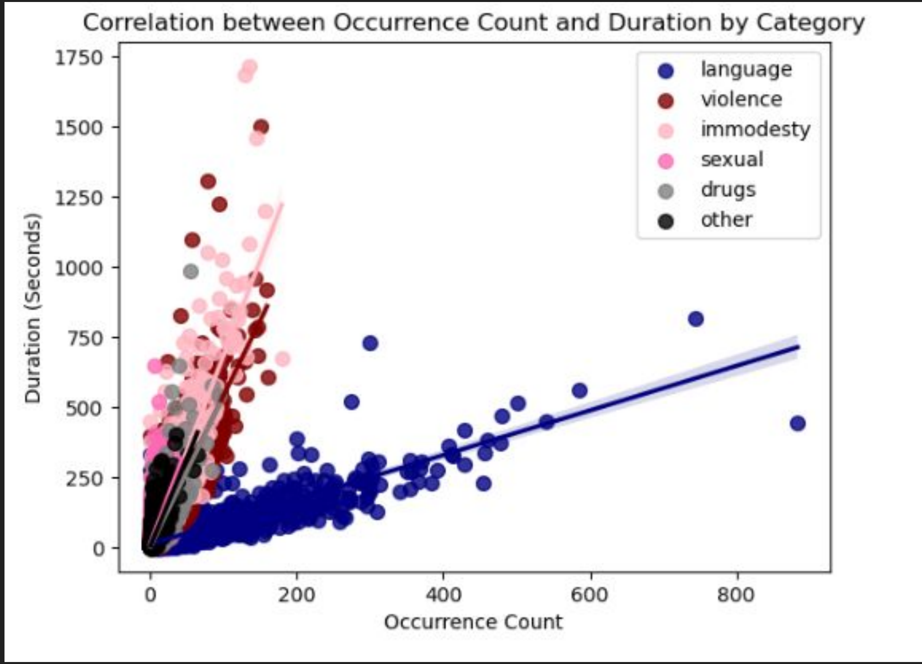
mpaa_rating category	G	NR	PG	PG-13	R	TV-14	TV-G	TV-MA	TV-PG	TV-Y7
drugs	15	27	272	808	470	90	6	100	38	7
immodesty	11	35	301	980	1670	99	7	278	36	10
language	28	74	686	1780	2811	193	22	459	93	22
other	16	42	357	962	1572	107	7	272	36	17
sexual	14	64	312	1320	2416	149	11	401	51	10
violence	30	56	492	1223	2082	107	13	338	55	20

EDA - Tag Name (Movie Tags)



category	tag_name \ mpaa_rating	G	NR	PG	PG-13	R	TV-14	TV-G	TV-MA	TV-PG	TV-Y7
drugs	drugs_illegal	0.0	8.0	29.0	125.0	311.0	14.0	0.0	63.0	2.0	2.0
	drugs_implied	7.0	8.0	126.0	263.0	149.0	32.0	2.0	32.0	17.0	4.0
	drugs_legal	8.0	11.0	117.0	420.0	10.0	44.0	4.0	5.0	19.0	1.0
immodesty	immodesty	5.0	18.0	153.0	452.0	664.0	48.0	6.0	113.0	23.0	5.0
	nudity_art	2.0	2.0	36.0	140.0	233.0	15.0	0.0	43.0	7.0	0.0
	nudity_implied	2.0	4.0	45.0	159.0	276.0	10.0	1.0	40.0	2.0	3.0
	nudity_without_sex	2.0	11.0	67.0	229.0	497.0	26.0	0.0	82.0	4.0	2.0
language	blasphemy	7.0	21.0	167.0	494.0	748.0	55.0	7.0	128.0	26.0	5.0
	childish_language	10.0	9.0	183.0	212.0	353.0	21.0	6.0	51.0	27.0	8.0
	profanity	6.0	21.0	150.0	502.0	759.0	54.0	6.0	129.0	20.0	5.0
	racial_slurs	1.0	5.0	35.0	144.0	284.0	16.0	0.0	40.0	1.0	0.0
	sexual_reference	4.0	18.0	151.0	428.0	667.0	47.0	3.0	111.0	19.0	4.0
other	bodily_functions	6.0	15.0	130.0	276.0	482.0	36.0	3.0	80.0	13.0	7.0
	life_events	1.0	4.0	36.0	87.0	127.0	11.0	1.0	22.0	3.0	1.0
	medical_graphic	0.0	5.0	13.0	80.0	181.0	7.0	0.0	33.0	3.0	0.0
	medical_procedures	2.0	2.0	25.0	98.0	130.0	12.0	0.0	27.0	5.0	0.0
	objectionable	5.0	11.0	113.0	246.0	349.0	23.0	3.0	60.0	10.0	7.0
	vulgar_gestures	2.0	5.0	40.0	175.0	303.0	18.0	0.0	50.0	2.0	2.0
sexual	kissing_normal	5.0	11.0	116.0	361.0	476.0	37.0	4.0	76.0	22.0	3.0
	kissing_passion	4.0	13.0	57.0	246.0	380.0	30.0	3.0	53.0	11.0	1.0
	sex_implied	0.0	10.0	16.0	215.0	405.0	27.0	2.0	66.0	1.0	0.0
	sex_with_nudity	0.0	4.0	0.0	17.0	171.0	3.0	0.0	32.0	0.0	0.0
	sex_without_nudity	0.0	4.0	5.0	56.0	236.0	7.0	0.0	35.0	1.0	1.0
	sexual_assault	1.0	9.0	12.0	107.0	267.0	13.0	0.0	61.0	2.0	1.0
	sexually_suggestive	4.0	13.0	106.0	318.0	481.0	32.0	2.0	78.0	14.0	4.0
violence	disturbing_images	5.0	10.0	74.0	249.0	440.0	16.0	1.0	68.0	5.0	2.0
	gore	1.0	7.0	13.0	77.0	229.0	7.0	1.0	42.0	1.0	1.0
	graphic	6.0	15.0	90.0	363.0	630.0	29.0	2.0	101.0	9.0	3.0
	non_graphic	10.0	20.0	184.0	487.0	722.0	49.0	6.0	119.0	27.0	8.0
	violence_implied	8.0	4.0	131.0	47.0	61.0	6.0	3.0	8.0	13.0	6.0

EDA (Movie Tags) - Occurrence Cnt & Duration (Sec.)



Preprocessing

- Removed NR movies
- Merged movies and movie_tags on imdb_id
  - Created 2 columns for each category/movie tag
    - Occurrence Count
    - Duration (seconds)
- Split out the studio column into multiple columns for each studio
- Created X and y variables
- Split X and y into 80% train and 20% test variables
- Scaled X values using X_train

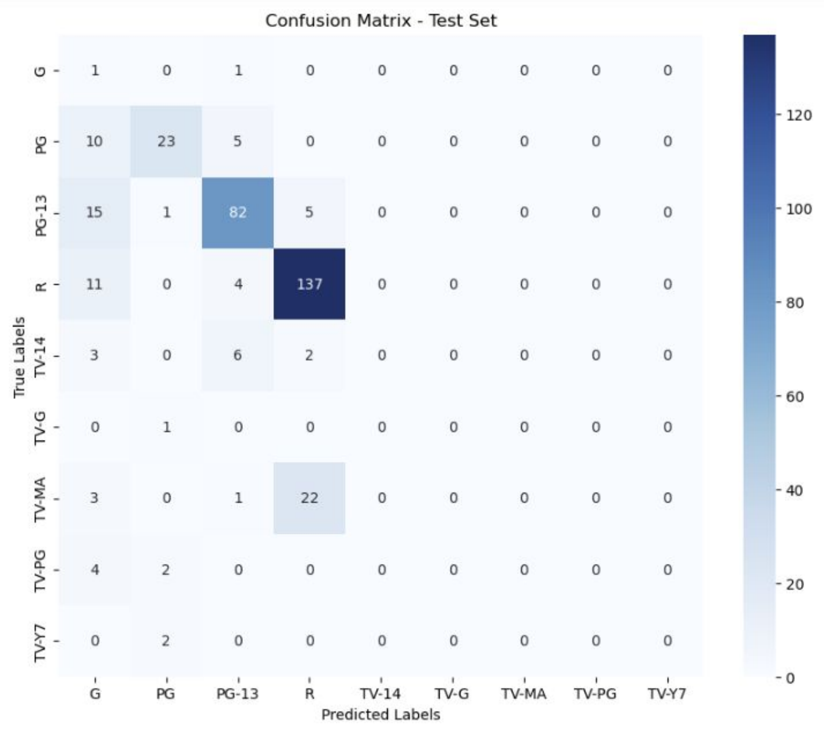
X_train

	year	duration_sec	number_of_studios	duration_sec_drugs_drugs_illegal	duration_sec_drugs_drugs_implied	duration_sec_drugs_drugs_legal
986	-0.956081	1.242184	-0.491462	-0.308978	-0.315367	-0.092523
1242	-1.781130	0.037575	-0.491462	-0.308978	-0.315367	-0.385055
615	-1.471737	-0.019787	-0.491462	-0.308978	0.844618	-0.238789
1416	0.384625	-0.019787	-0.491462	-0.308978	-0.315367	-0.385055
1467	0.694019	-0.536048	-0.491462	-0.308978	-0.315367	-0.385055
...	...	...	...	...	...	...
1507	-0.234162	0.209662	-0.491462	-0.308978	-0.315367	0.346275
1223	0.384625	-0.650773	-0.491462	-0.305720	2.906814	-0.238789
77	-1.162343	0.898010	-0.491462	-0.308978	-0.057593	0.492541
1341	0.487756	0.496474	-0.491462	0.081962	-0.186480	-0.385055
632	0.075231	-0.363961	-0.491462	-0.308978	-0.315367	-0.385055

y_train

	G	PG	PG-13	R	TV-14	TV-G	TV-MA	TV-PG	TV-Y7
986	0	0	1	0	0	0	0	0	0
1242	0	0	0	1	0	0	0	0	0
615	0	1	0	0	0	0	0	0	0
1416	0	0	0	1	0	0	0	0	0
1467	0	0	0	1	0	0	0	0	0
...	...	...	...	...	...	...	...	...	...
1507	0	0	1	0	0	0	0	0	0
1223	0	0	1	0	0	0	0	0	0
77	0	0	1	0	0	0	0	0	0
1341	0	0	0	1	0	0	0	0	0
632	0	1	0	0	0	0	0	0	0

Models - Random Forest

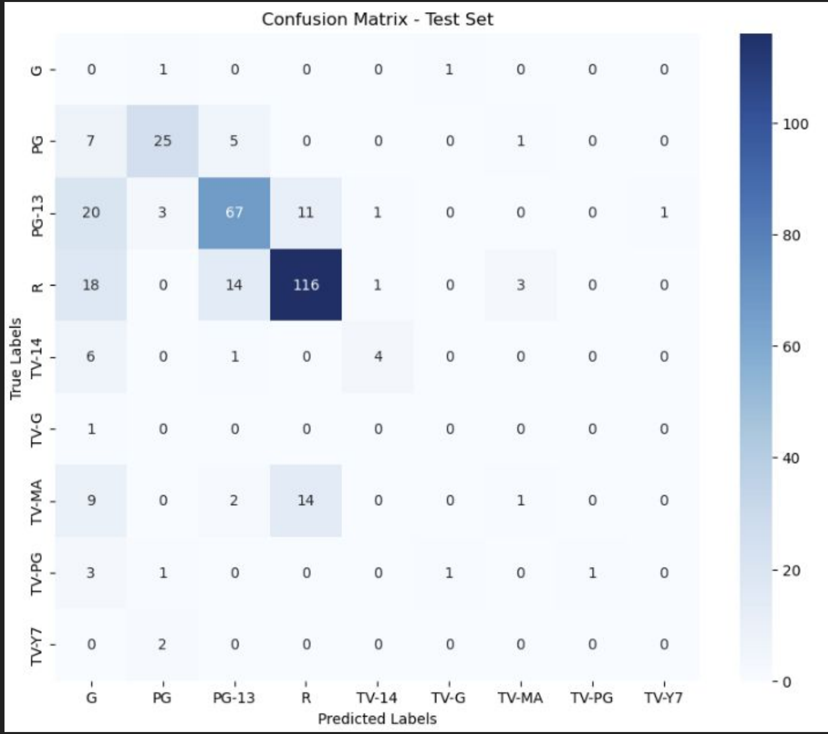



```
Accuracy: 0.71
```

Models - Support Vector Machine

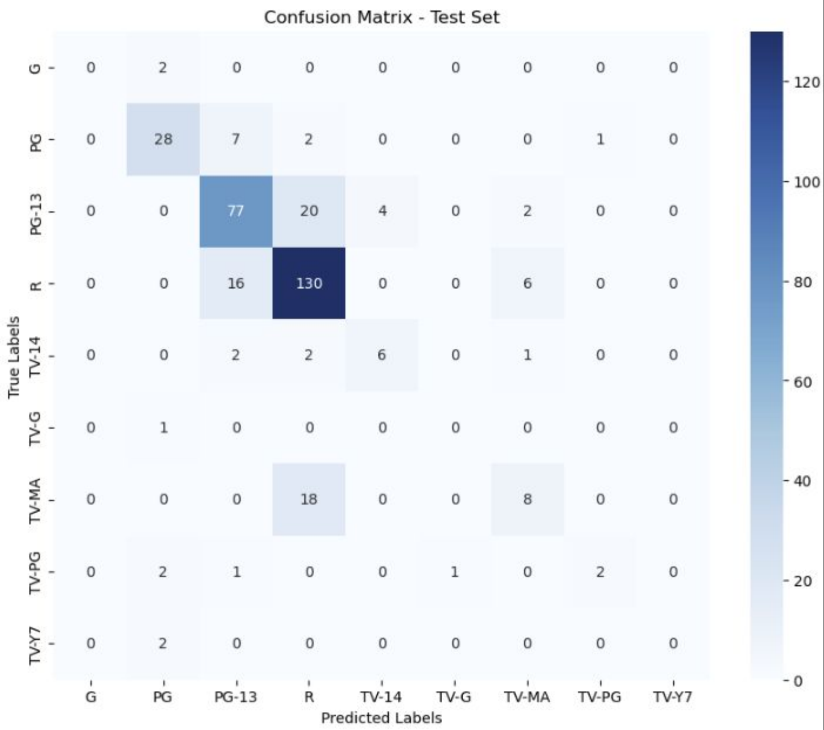
Accuracy: 0.63

Models - Logistic Regression



Accuracy: 0.63

Models - Deep Learning



`Accuracy: 0.74`

Conclusion

- R is the most common rating
- Movie data is skewed to more recently made movies (2010+)
- Number of studios does not appear to be impactful
- Tags Rating:
  - The legal drugs tag name is very low in R movies.
  - The illegal drugs tag name is very high in R movies.
- Tags Occurrence and Duration:
  - Language tags have high occurrences, but low duration.
  - All other tags have low occurrences, but high duration.
- Best Model:
  - Deep Learning was the best model due to more consistent predictions

Next Steps

- Short Term
  - Take the NR movies and see what our model predicts.
  - Gather some more data to be able to better predict on some of the other ratings.
  - Create more features.
- Long Term
  - Dive deeper into why G was over predicting.
  - Look into the Deep Learning model further to see what factors are driving it's predictions.
  - Further classify the ratings into recommended age groups.
  - Develop an app connected to an API for movie information for looking up movies with ease.

Limitations

- Limited data for certain ratings affecting predictions.
- Categories that are similar (Ex: PG and TV-PG) can create confusion in the model.
- Manual mapping of studios may not be fully accurate.