

OTT서비스 콘텐츠 분석

Team 김밥초밥비빔밥
(위나라 조조 하후돈 조인)

팀장 박광석
맹성주

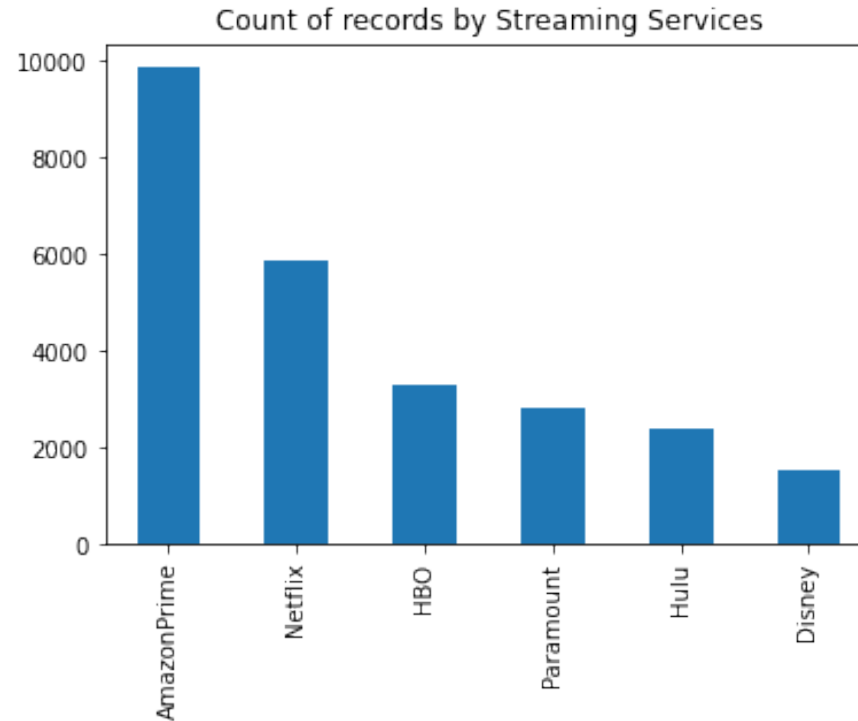
Speaker. 김정대

Contents

- About the dataset
- Preprocessing process
- Recent Trend(최근 5년) in the dataset
- Scenario. EDA by platform
- Application example

About The Dataset

- 총 6개의 OTT(Over The Top)서비스들 (Amazon Prime, Netflix, Paramount, Disney, Hulu, HBO)의 작품(movie, TV show) 데이터셋



- 24,000개의 작품들, 총 15개의 column으로 구성

	id	title	type	description	release_year	age_certification	runtime	genres	production_c	seasons	imdb_id	imdb_score	imdb_votes	tmdb_pop	tmdb_s
0	ts20945	The Three Stooges	SHOW	The Three Stooges were an Am	1934	TV-PG	19	['comedy', 'f']	['US']	26	tt085064	8.6	1092	15.424	7.6
1	tm19248	The General	MOVIE	During America's Civil War, Un	1926		78	['action', 'dra']	['US']		tt001792	8.2	89766	8.647	8
2	tm82253	The Best Years of Our Lives	MOVIE	It's the hope that sustains the s	1946		171	['romance', '']	['US']		tt003686	8.1	63026	8.435	7.8
3	tm83884	His Girl Friday	MOVIE	Hildy, the journalist former wife	1940		92	['comedy', 'c']	['US']		tt003259	7.8	57835	11.27	7.4
4	tm56584	In a Lonely Place	MOVIE	An aspiring actress begins to su	1950		94	['thriller', 'dra']	['US']		tt004259	7.9	30924	8.273	7.6
5	tm160494	Stagecoach	MOVIE	A group of people traveling on	1939		96	['western', 'd']	['US']		tt003197	7.8	48149	11.786	7.7
6	tm87233	It's a Wonderful Life	MOVIE	A holiday favourite for generati	1946	PG	130	['drama', 'far']	['US']		tt003865	8.6	444243	26.495	8.3
7	tm19424	Detour	MOVIE	The life of Al Roberts, a pianist i	1945		66	['thriller', 'dra']	['US']		tt003763	7.3	17233	7.757	7.2
8	tm116781	My Man Godfrey	MOVIE	Fifth Avenue socialite Irene Bul	1936		95	['comedy', 'r']	['US']		tt002801	8	23532	8.633	7.6
9	tm112005	Marihuana	MOVIE	A young girl named Burma atte	1936		57	['crime', 'dra']	['US']		tt002668	4	864	3.748	3.6
10	tm22806	Intolerance: Love's Struggle Throug	MOVIE	The story of a poor young worr	1916		197	['history', 'dr']	['US']		tt000686	7.7	15242	9.412	7.2
11	tm88001	The Stranger	MOVIE	A man working for the War Crim	1946		95	['thriller', 'cri']	['US']		tt003899	7.3	26293	7.566	7.3
12	tm141715	The Phantom of the Opera	MOVIE	A grotesquely disfigured compo	1925		101	['horror']	['US']		tt001622	7.5	18190	7.428	7.2
13	tm112415	Tell Your Children	MOVIE	High-school principal Dr. Alfred	1938	PG	66	['drama', 'cri']	['US']		tt002834	3.7	8617	5.412	4.1
14	tm2838	The Most Dangerous Game	MOVIE	When legendary hunter Bob Rai	1932		63	['thriller', 'ho']	['US']		tt002323	7.1	12203	8.828	7
15	tm155610	Wuthering Heights	MOVIE	The Earnshaws are Yorkshire far	1939		104	['drama', 'ror']	['US']		tt003214	7.5	17956	12.336	7.3
16	tm3781	The Woman in the Window	MOVIE	A seductive woman gets an inn	1944		99	['drama', 'thr']	['US']		tt003746	7.6	15909	6.018	7.5
17	tm3541	Steamboat Bill, Jr.	MOVIE	The just out of college effete sc	1928		70	['drama', 'ror']	['US']		tt001942	7.8	14499	7.754	7.7
18	tm82560	The Little Foxes	MOVIE	The ruthless, moneyed Hubbarc	1941		115	['romance', '']	['US']		tt003383	7.9	11775	5.906	7.5
19	tm64852	Stella Dallas	MOVIE	After divorcing a society man, a	1937		106	['drama', 'ror']	['US']		tt002960	7.4	5440	5.377	6.9
20	tm120863	Dodsworth	MOVIE	A retired auto manufacturer anc	1936		101	['romance', '']	['US']		tt002753	7.8	9063	6.1	7.2
21	tm66903	Sands of Iwo Jima	MOVIE	The relationship between Serge	1950		100	['drama', 'wa']	['US']		tt004184	7	10399	7.629	6.5

Preprocessing Process

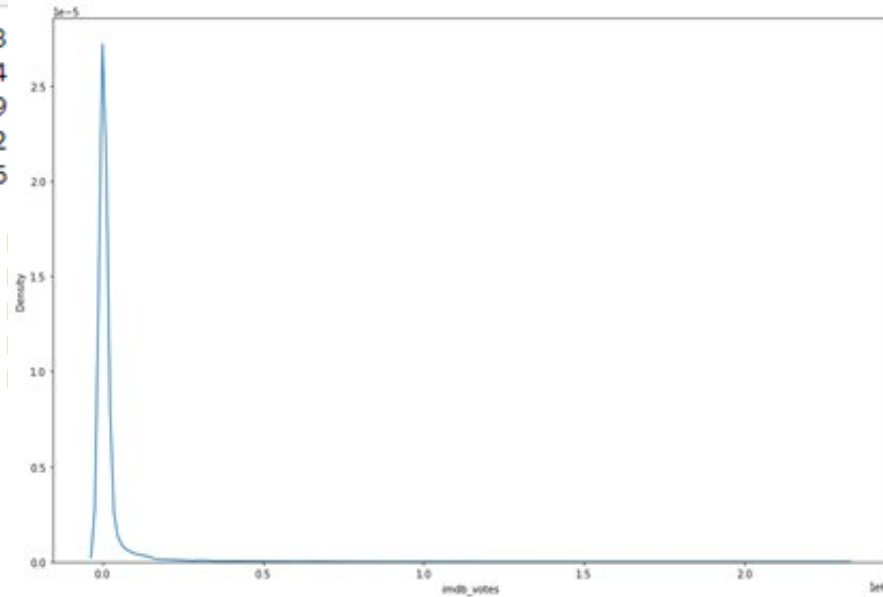
Preprocessing process

- 결측치 제거 by column
 1. “imdb_score”, “imdb_votes”, “tmdb_score”, “tmdb_popularity” column 모두 결측인 것 -> 제거
 2. “age_certification” column이 결측치인 것 -> 값을 “not rated”로 변경
 3. “season” column이 결측치인 것 -> 값을 ‘0’으로 변경
 4. IMDb와 TMDB의 정보가 하나라도 있는 것 -> 각 DB 정보 칼럼의 평균으로 변경
- 중복처리 -> imdb_id를 기준으로 중복된 작품들 제거

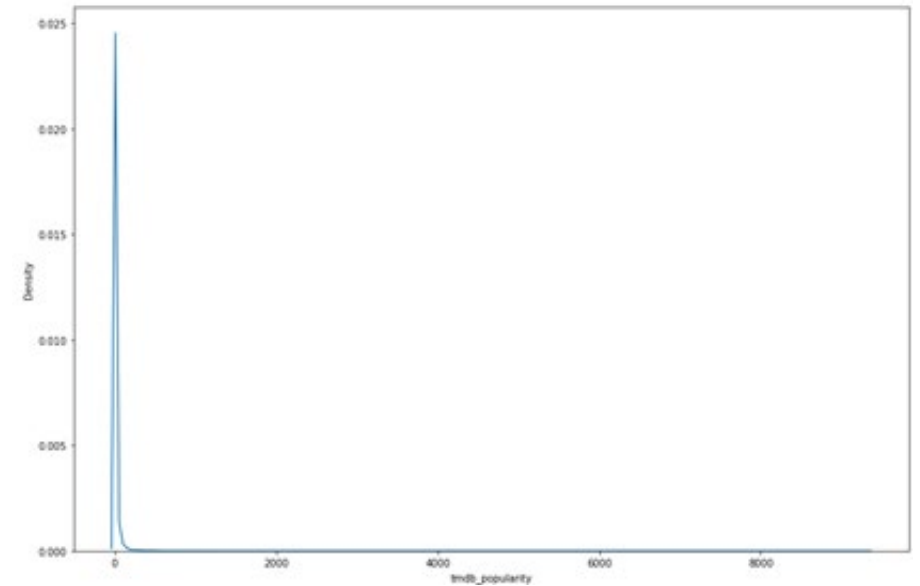
Preprocessing process

```
df['imdb_votes'].sort_values(ascending = False)
```

11643	229423
11503	202134
18333	199459
18329	198272
17805	178056
...	
24854	
24870	
24871	
24873	
24875	



imdb_votes's mean = 20990



tmdb_popularity's mean = 17.58

Preprocessing process

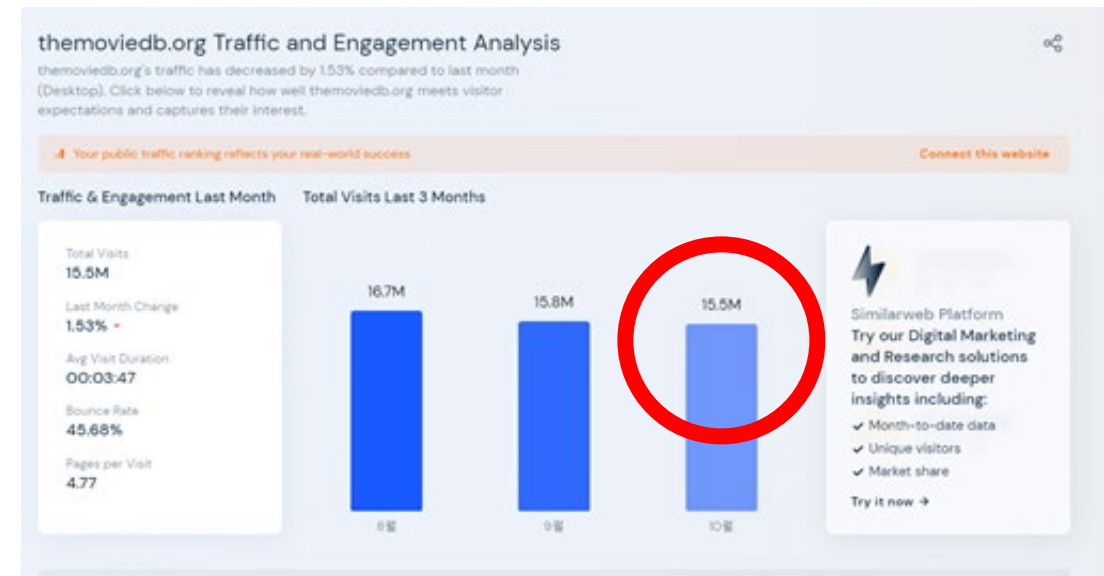
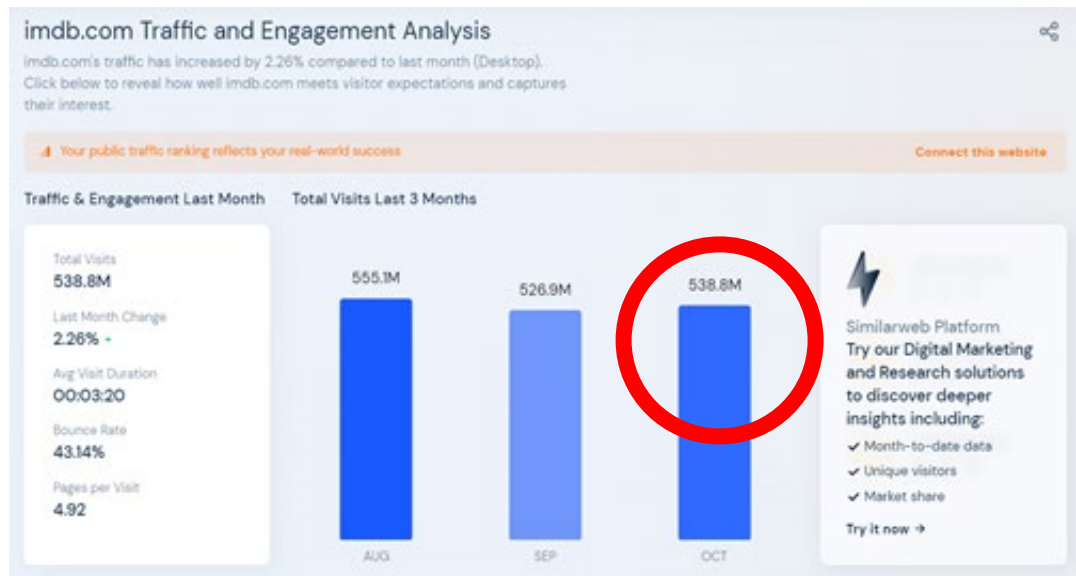
- 'imdb_popularity' column 생성

$$\textit{Weighted Rating}(WR) = \left(\frac{v}{v+m} \times R \right) + \left(\frac{m}{v+m} \times C \right)$$

- v is the number of votes garnered by the movie
- m is the minimum number of votes required for the movie to be in the chart (the prerequisite)
- R is the mean rating of the movie
- C is the mean rating of all the movies in the dataset

Preprocessing process

- 'total_popularity' column 생성 -> IMDb와 TMDB 웹사이트의 방문자 수를 근거로 가중치 설정
- 사이트 방문자 수 조사 -> 방문자 수 비율 538.8 : 15.5

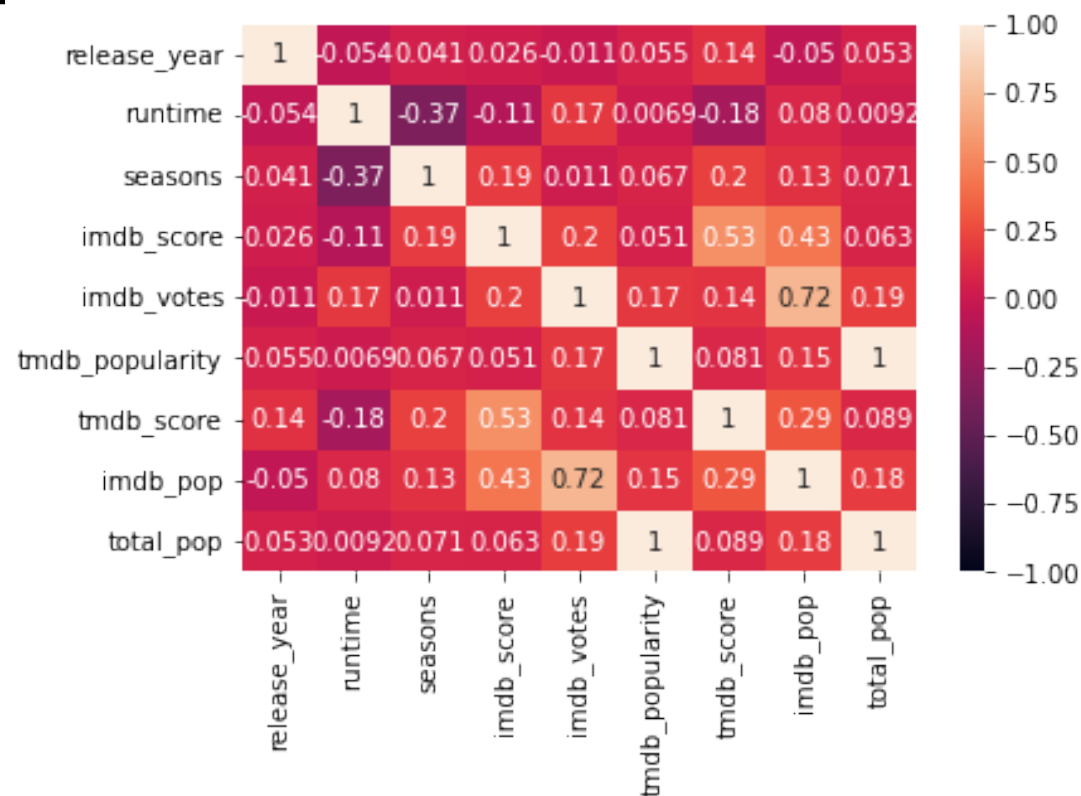


Preprocessing process

id	title	type	description	release	age_certificate	runtime	genres	production_season	imdb_id	imdb_rating	imdb_votes	tmdb_popularity	imdb_platform	imdb_popularity	total_popularity
ts20945	The Three Stooges	SHOW	The Three Stooges were an American vaudeville comedy team that consisted of three members: Moe, Larry, and Curly Howard.	1934	TV-PG	19	['comedy', 'family', 'US']	26	tt0850645	8.6	1092	15.424	7.6 Disney	6.277133	-0.0016
tm19248	The General	MOVIE	During America's Civil War, Union spies steal a general's train.	1926	Not Rated	78	['action', 'drama', 'US']	0	tt0017925	8.2	89766	8.647	8.7 AmazonPrime	7.325419	0.003576
tm82253	The Best Years of Our Lives	MOVIE	It's the hope that sustains the spirit of every man, woman, and child.	1946	Not Rated	171	['romance', 'war', 'US']	0	tt0036868	8.1	63026	8.435	7.8 AmazonPrime	7.106309	0.022777
tm83884	His Girl Friday	MOVIE	Hildy, the journalist former wife of newspaper editor Walter, returns to work.	1940	Not Rated	92	['comedy', 'drama', 'US']	0	tt0032599	7.8	57835	11.27	7.4 Netflix	6.933573	0.031201
tm56584	In a Lonely Place	MOVIE	An aspiring actress begins to suspect that her husband is hiding a dark secret.	1950	Not Rated	94	['thriller', 'drama', 'US']	0	tt0042593	7.9	30924	8.273	7.6 HBOHulu	6.737974	-0.02003
tm160494	Stagecoach	MOVIE	A group of people traveling on a stagecoach during the American frontier.	1939	Not Rated	96	['western', 'drama', 'US']	0	tt0031971	7.8	48149	11.786	7.7 Disney	6.863997	0.028466
tm87233	It's a Wonderful Life	MOVIE	A holiday favourite for generations. George Bailey is on the verge of losing his job.	1946	PG	130	['drama', 'family', 'US']	0	tt0038650	8.6	444243	26.495	8.3 Netflix	8.269298	0.329763
tm19424	Detour	MOVIE	The life of Al Roberts, a pianist in a New York City nightclub.	1945	Not Rated	66	['thriller', 'drama', 'US']	0	tt0037638	7.3	17233	7.757	7.2 Netflix	6.445357	-0.05784
tm116781	My Man Godfrey	MOVIE	Fifth Avenue socialite Irene Bullock needs a butler.	1936	Not Rated	95	['comedy', 'romance', 'US']	0	tt0028010	8	23532	8.633	7.6 Netflix	6.673013	-0.02378
tm112005	Marihuana	MOVIE	A young girl named Burma attends a beach party.	1936	Not Rated	57	['crime', 'drama', 'US']	0	tt0026683	4	864	3.748	3.6 AmazonPrime	6.215717	-0.12279
tm22806	Intolerance: Love's Struggle Throughout the Ages	MOVIE	The story of a poor young woman, separated from her lover.	1916	Not Rated	197	['history', 'drama', 'US']	0	tt0006864	7.7	15242	9.412	7.2 Netflix	6.495466	-0.03602
tm88001	The Stranger	MOVIE	A man working for the War Crimes Commission is accused of a crime.	1946	Not Rated	95	['thriller', 'crime', 'US']	0	tt0038991	7.3	26293	7.566	7.3 AmazonPrime	6.523757	-0.05093
tm141715	The Phantom of the Opera	MOVIE	A grotesquely disfigured composer known only as 'The Phantom' haunts the Paris Opera House.	1925	Not Rated	101	['horror', 'US']	0	tt0016220	7.5	18190	7.428	7.2 AmazonPrime	6.494506	-0.05556
tm112415	Tell Your Children	MOVIE	High-school principal Dr. Alfred Carroll relates a story to his class.	1938	PG	66	['drama', 'crime', 'US']	0	tt0028346	3.7	8617	5.412	4.1 Netflix	5.971989	-0.13378
tm2838	The Most Dangerous Game	MOVIE	When legendary hunter Bob Rainsford is shot, he is forced to play a deadly game.	1932	Not Rated	63	['thriller', 'horror', 'US']	0	tt0023238	7.1	12203	8.828	7 AmazonPrime	6.365765	-0.05626
tm155610	Wuthering Heights	MOVIE	The Earnshaws are Yorkshire farmers during the 19th century.	1939	Not Rated	104	['drama', 'romance', 'US']	0	tt0032145	7.5	17956	12.336	7.3 Netflix	6.491904	-0.00779
tm3781	The Woman in the Window	MOVIE	A seductive woman gets an innocent professor.	1944	Not Rated	99	['drama', 'thriller', 'US']	0	tt0037469	7.6	15909	6.018	7.5 AmazonPrime	6.486556	-0.07025
tm3541	Steamboat Bill, Jr.	MOVIE	The just out of college effete son of a no-nonsense father.	1928	Not Rated	70	['drama', 'romance', 'US']	0	tt0019421	7.8	14499	7.754	7.7 Paramount	6.501847	-0.05154
tm82560	The Little Foxes	MOVIE	The ruthless, moneyed Hubbard clan lives in the South.	1941	Not Rated	115	['romance', 'drama', 'US']	0	tt0033836	7.9	11775	5.906	7.5 AmazonPrime	6.473849	-0.07277
tm64852	Stella Dallas	MOVIE	After divorcing a society man, a small-town woman moves to the city.	1937	Not Rated	106	['drama', 'romance', 'US']	0	tt0029608	7.4	5440	5.377	6.9 Netflix	6.322978	-0.09484
tm120863	Dodsworth	MOVIE	A retired auto manufacturer and his wife take a road trip.	1936	Not Rated	101	['romance', 'drama', 'US']	0	tt0027532	7.8	9063	6.1	7.2 Netflix	6.415297	-0.07742
tm66903	Sands of Iwo Jima	MOVIE	The relationship between Sergeant Stryker and Private Pyle.	1950	Not Rated	100	['drama', 'war', 'US']	0	tt0041841	7	10399	7.629	6.5 HBOHulu	6.337233	-0.07119

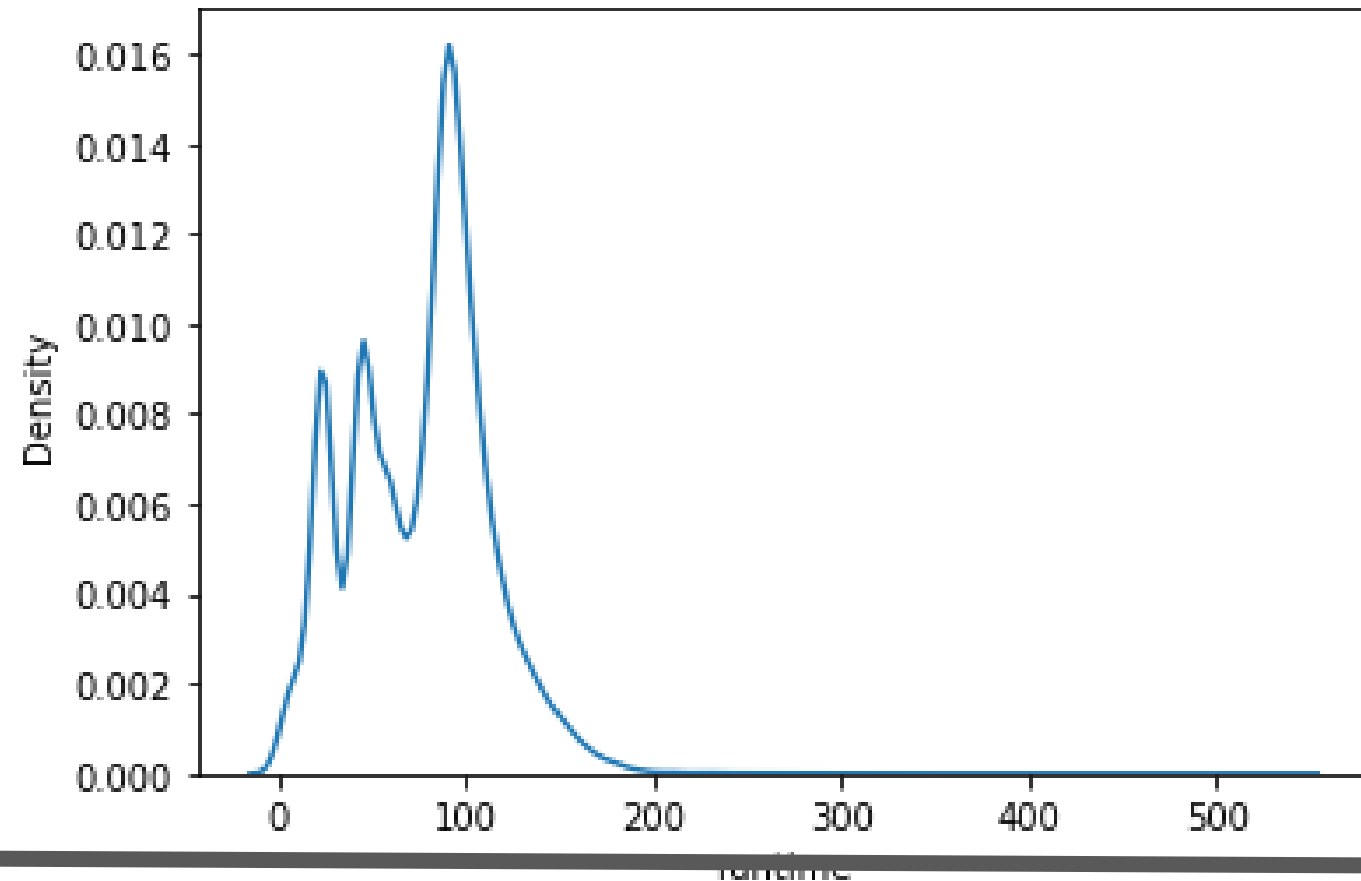
Preprocessing process

- 상관관계 히트맵



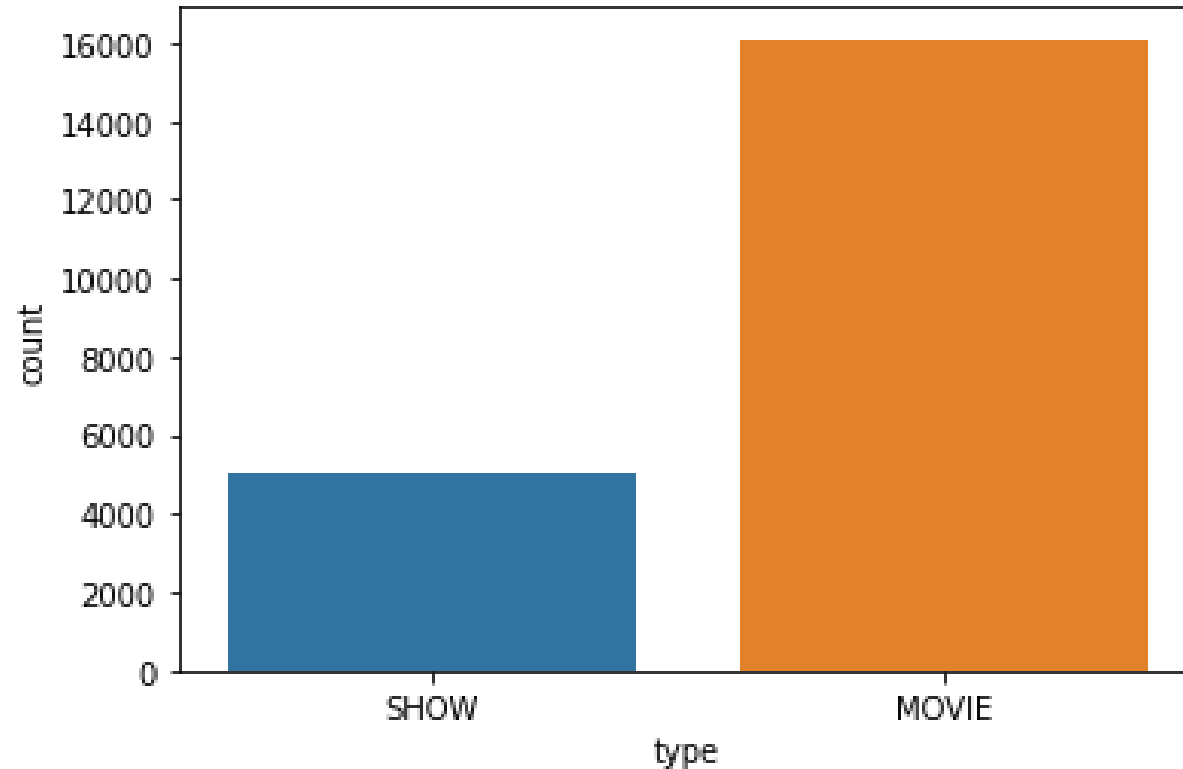
Preprocessing process

- Runtime



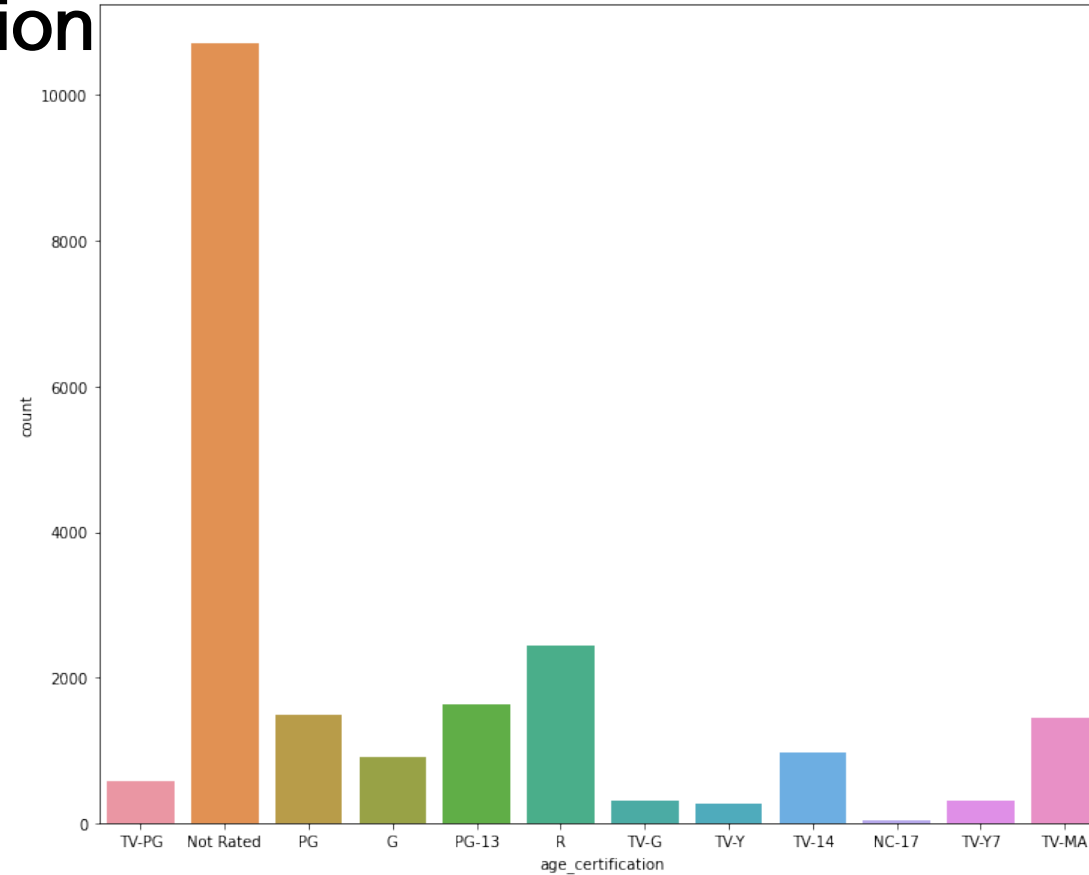
Preprocessing process

- Type



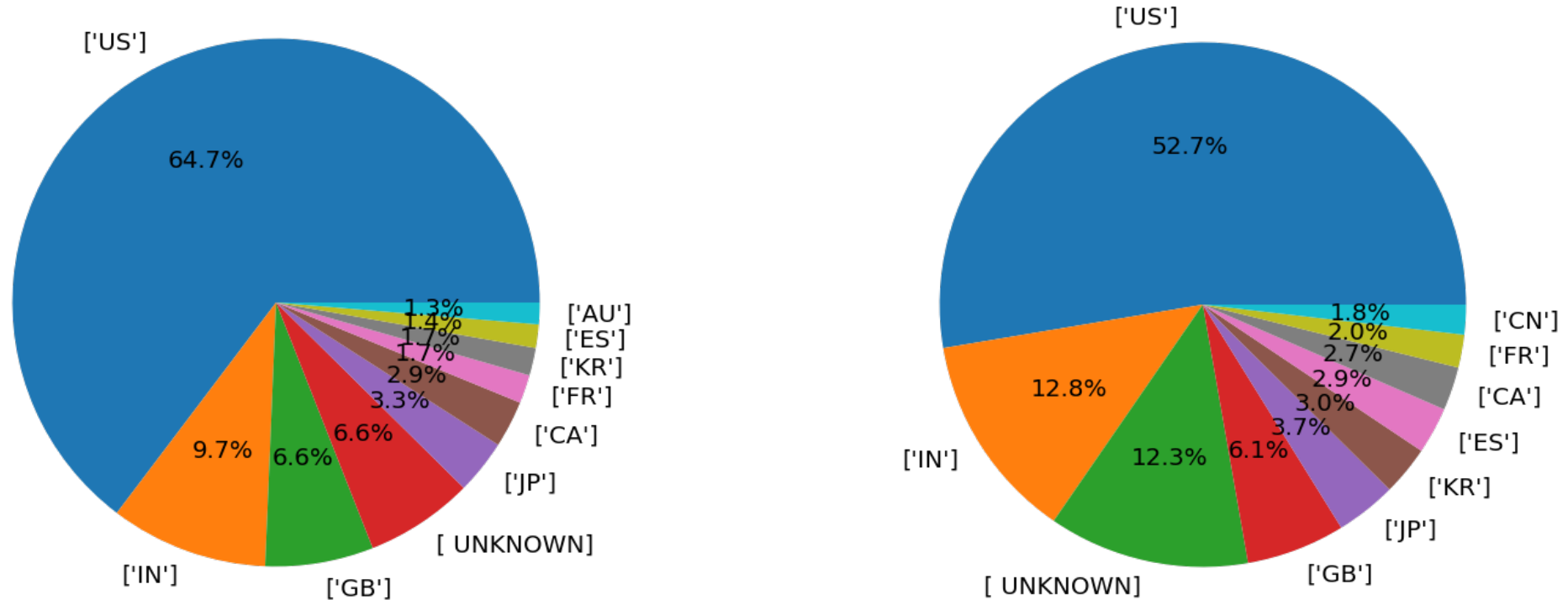
Preprocessing process

- Age certification



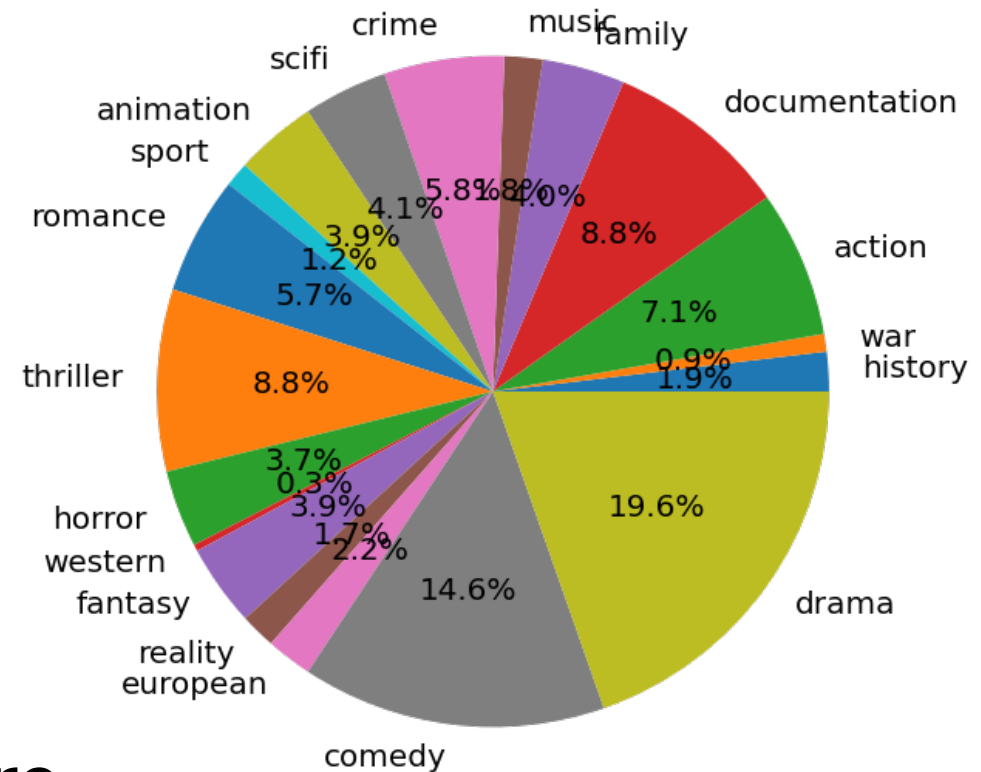
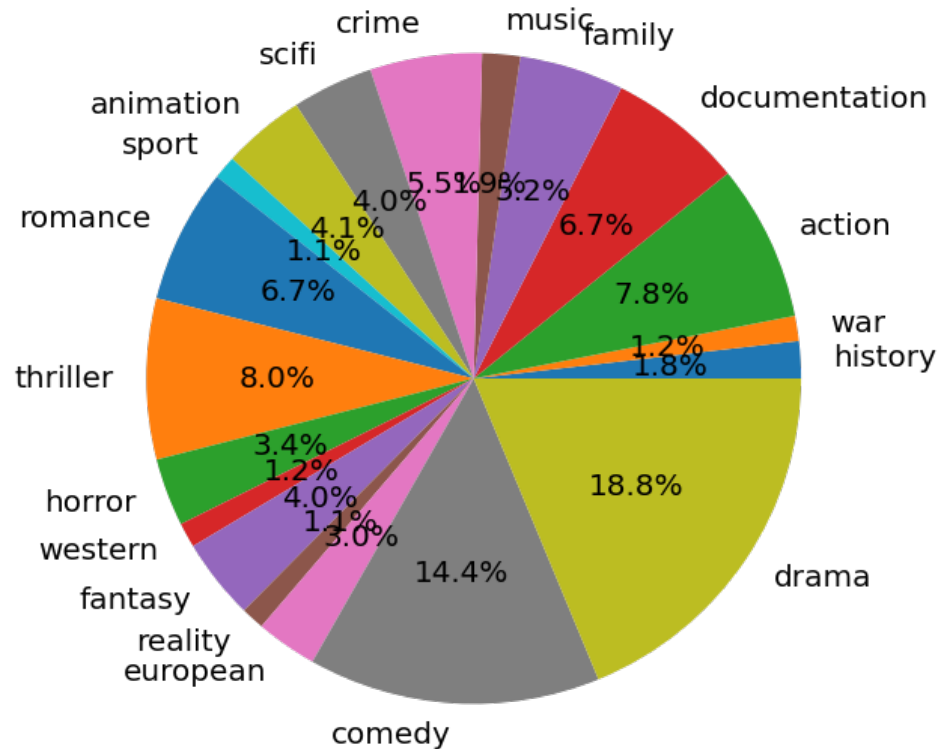
Recent Trend(최근 5년) In The Dataset

Trend In The Dataset



Production_Country
전체 vs 최근 5년

Trend In The Dataset



Genre
전체 vs 최근 5년

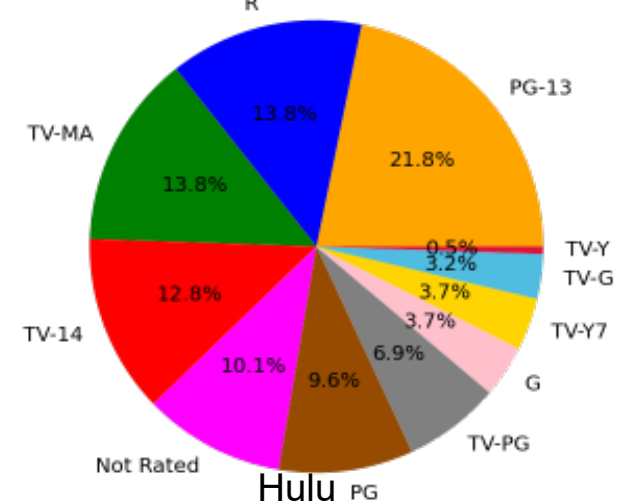
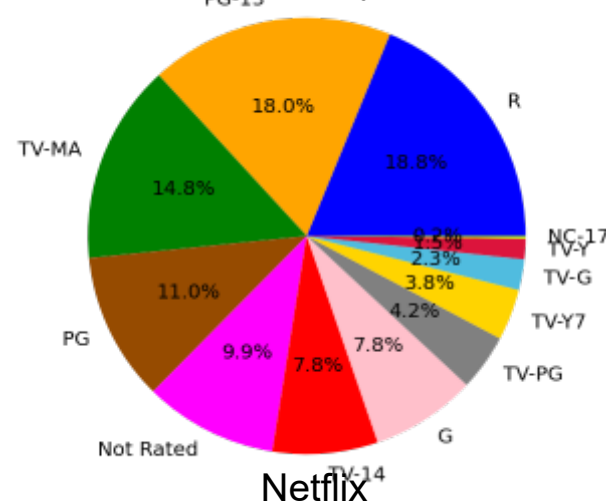
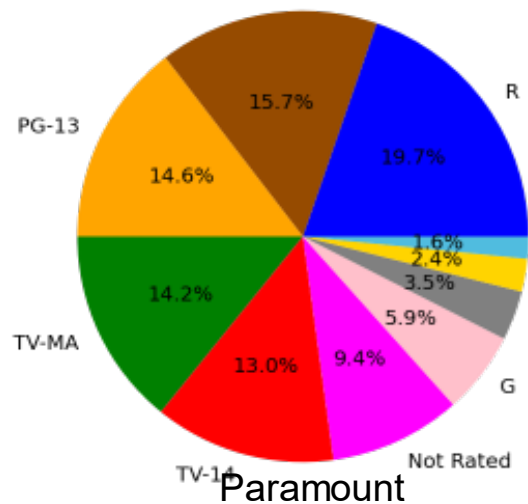
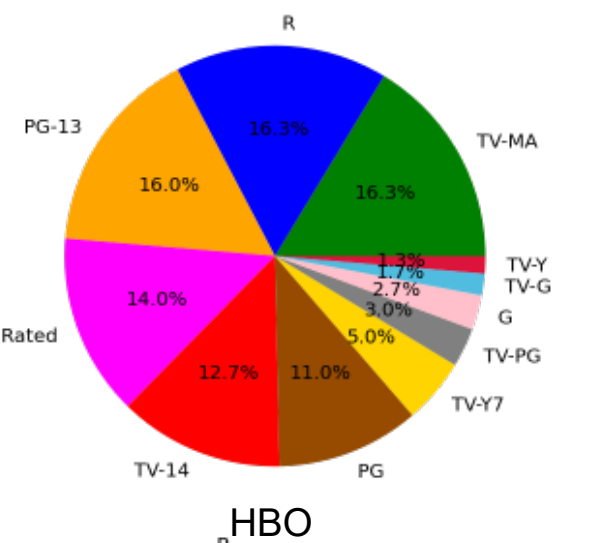
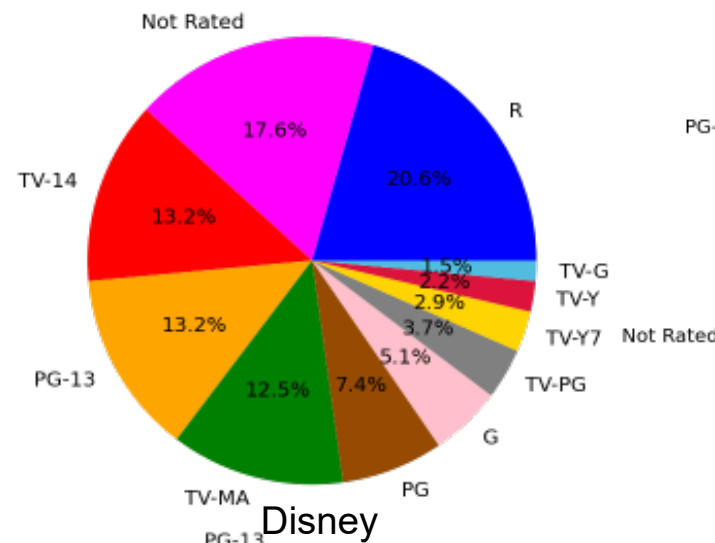
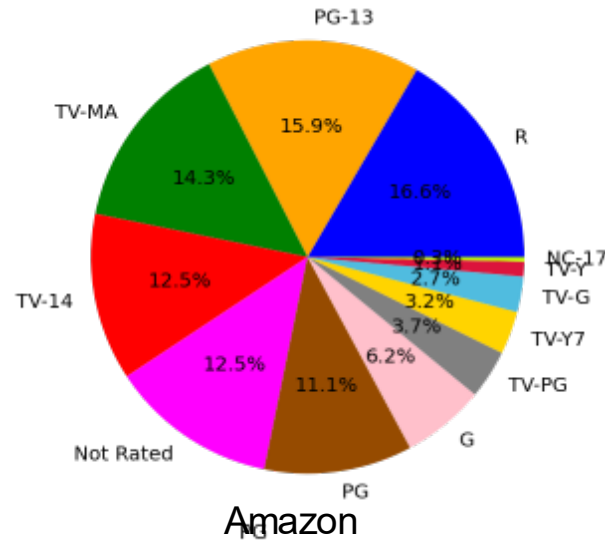
Scenario Data Analysis

Scenario. EDA by platform

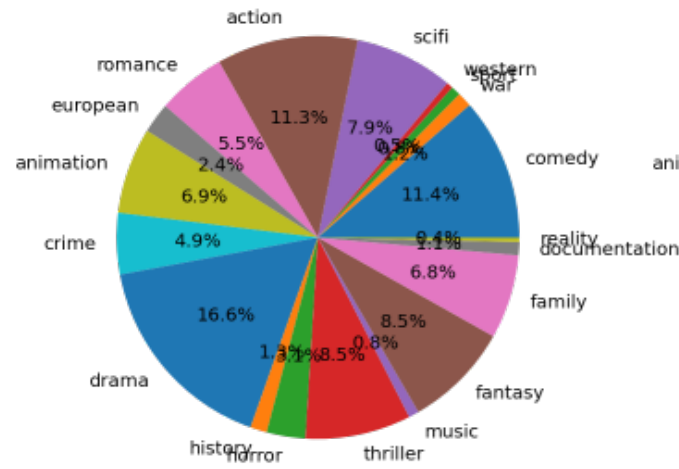
- 영화 제작업체가 OTT서비스에 영화를 서비스하려 한다. 어떤 platform에 어떤 작품을 추천할 수 있을까? 또, OTT서비스를 구독하려고 하는데, OTT플랫폼 중 구독하기 좋은 플랫폼은 어디일까?

-> IMDb와 TMDB의 상위 10% 작품들의 platform별 데이터 분석

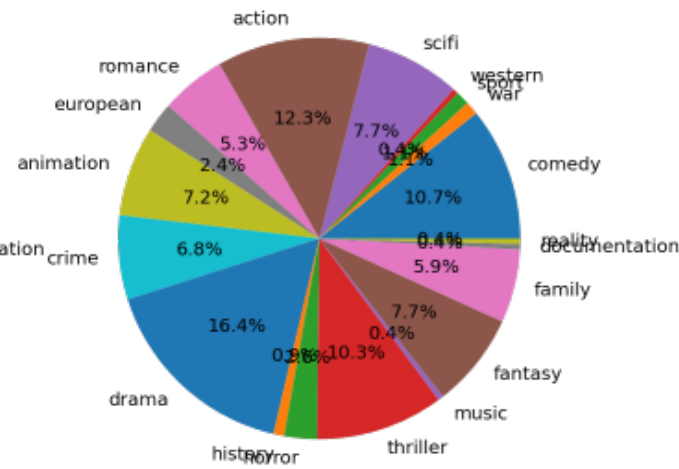
Scenario. EDA by platform



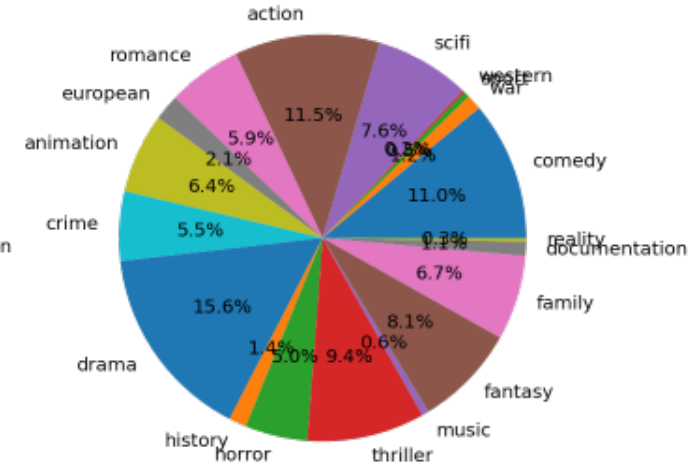
Scenario. EDA by platform



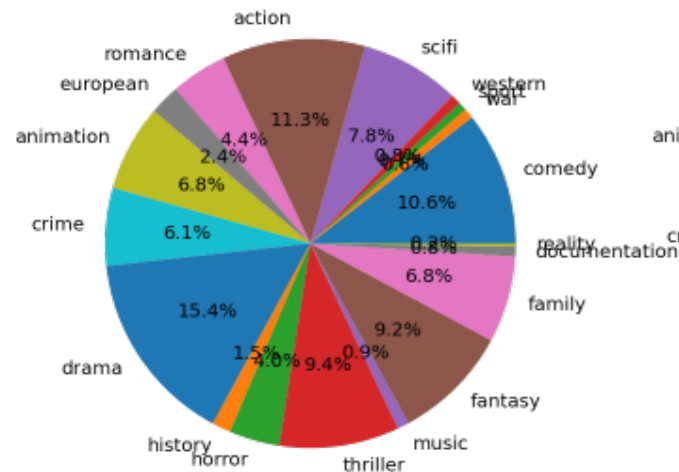
Amazon



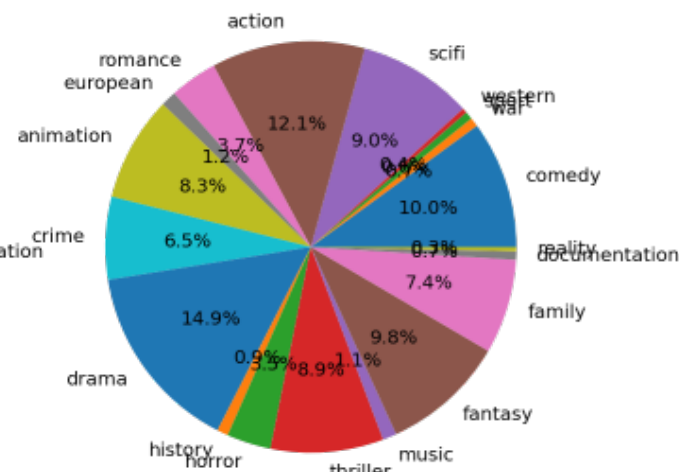
Disney



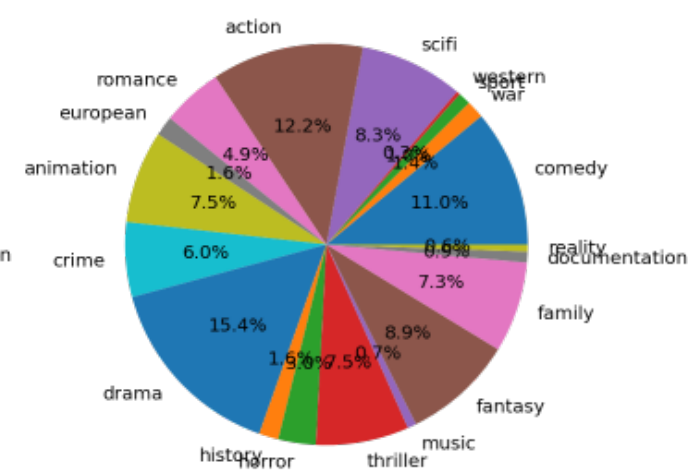
HBO



Paramount



Netflix



Hulu

Application Example

Application Example

- 머신러닝을 이용한 Recommendation system 구현 가능

출처

- <https://tutorialedge.net/python/building-imdb-top-250-clone-pandas/> - weighted rating
- <https://www.similarweb.com/> - IMDb와 TMDB 웹사이트 방문자 수 집계