MovieLens의 메타데이터를 이용한 추천 시스템

EDA

데이터 세트 목록

- Movies_metadata
- credits
- keywords
- links
- links_small
- ratings
- ratings_small

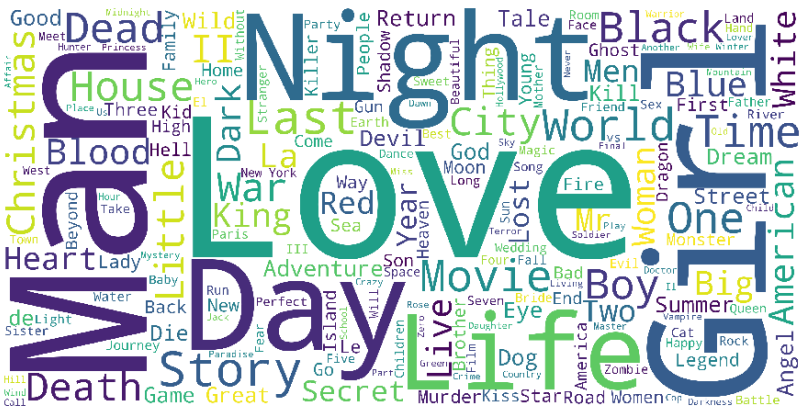
Movies_metadata (45466 rows, 24 cols)

adult	관람 등급 (T/F)
belongs_to_collection	영화가 속한 시리즈에 대한 정보 (문자열 딕셔너리)
Budget	예산
genres	관련된 모든 장르를 나열한 문자열 리스트/딕셔너리
homepage	공식 홈페이지 주소
id	
imdb_id	
original_language	촬영 시 언어
original_title	번역 또는 개작하기 전의 영화 제목
overview	영화에 대한 간단한 설명
popularity	TMDB가 지정한 인기도 점수
poster_path	포스터 이미지의 URL

Movies_metadata

production_company	영화 제작에 참여한 제작사 리스트
production_countries	영화가 촬영되거나 제작된 국가 리스트
release_date	극장 발매일
revene	총 수익
runtime	분 단위 런타임
spoken_languages	사용된 언어에 대한 리스트
status	출시 여부 (출시/출시예정/발표 등)
tagline	
title	공식 제목
video	TMDb에 영화의 비디오가 있는지 여부
vote_average	평균 평점
vote_count	투표 수

Title and Overview Wordclouds



TITLE

OVERVIEW

- 영화 제목 및 overview에 가장 많이 나타나는 단어는 'Love' , 'Man', 'Girl', 'Day', 'life', 'find' 등
- 로맨스나 드라마 장르 영화의 데이터가 많을 것으로 예상

Franchise Movies

시리즈 개수가 많은 순위

	belongs_to_collection	count	mean	sum
646	James Bond Collection	26	2.733450e+08	7.106970e+09
473	Friday the 13th Collection	12	3.874155e+07	4.648985e+08
976	Pokémon Collection	11	6.348189e+07	6.983008e+08
552	Harry Potter Collection	8	9.634209e+08	7.707367e+09
540	Halloween Collection	8	3.089601e+07	2.471681e+08
29	A Nightmare on Elm Street Collection	8	4.544894e+07	3.635916e+08
1317	The Fast and the Furious Collection	8	6.406373e+08	5.125099e+09
1432	The Pink Panther (Original) Collection	8	2.055978e+07	1.644782e+08
1160	Star Wars Collection	8	9.293118e+08	7.434495e+09
977	Police Academy Collection	7	4.352046e+07	3.046432e+08

- 속편이 가장 많은 시리즈
- 개수가 많다고 해서 성공적인 영화 시리즈를 의미하는 것은 아님
- James Bond는 총 25개로 가장 많은 시리즈 보유

Franchise Movies

총 수익 순위

	belongs_to_collection	count	mean	sum
552	Harry Potter Collection	8	9.634209e+08	7.707367e+09
1160	Star Wars Collection	8	9.293118e+08	7.434495e+09
646	James Bond Collection	26	2.733450e+08	7.106970e+09
1317	The Fast and the Furious Collection	8	6.406373e+08	5.125099e+09
968	Pirates of the Caribbean Collection	5	9.043154e+08	4.521577e+09
1550	Transformers Collection	5	8.732202e+08	4.366101e+09
325	Despicable Me Collection	4	9.227676e+08	3.691070e+09
1491	The Twilight Collection	5	6.684215e+08	3.342107e+09
610	Ice Age Collection	5	6.433417e+08	3.216709e+09
666	Jurassic Park Collection	4	7.578710e+08	3.031484e+09

- 최고 수익 영화 시리즈는 총 77억 달러 이상으로 해리포터
- 스타워즈 시리즈는 74억 3000만 달러로 두번째 고수익을 얻은 시리즈
- 제임스 본드 시리즈는 3위이지만, 다른 영화들과 비교하면 시리즈 수가 많기 때문에 훨씬 적은 수익이라고 할 수 있다

Franchise Movies

평균 수익 순위

	belongs_to_collection	count	mean	sum
112	Avatar Collection	1	2.787965e+09	2.787965e+09
1245	The Avengers Collection	2	1.462481e+09	2.924962e+09
479	Frozen Collection	1	1.274219e+09	1.274219e+09
446	Finding Nemo Collection	2	9.844532e+08	1.968906e+09
1352	The Hobbit Collection	3	9.785078e+08	2.935523e+09
1388	The Lord of the Rings Collection	3	9.721816e+08	2.916545e+09
552	Harry Potter Collection	8	9.634209e+08	7.707367e+09
1160	Star Wars Collection	8	9.293118e+08	7.434495e+09
325	Despicable Me Collection	4	9.227676e+08	3.691070e+09
968	Pirates of the Caribbean Collection	5	9.043154e+08	4.521577e+09

- 물가에 맞춰 스케일이 조정되지 않았기 때문에 적합한 지표는 아님
- 아바타는 단 한편으로 30억 달러에 가까운 수익을 벌어들임
- 해리포터는 5편 이상의 시리즈물 중에서 가장 성공적인 시리즈라고 할 수 있다

Production Companies

높은 수익을 얻은 제작사 순위

	Total	Average	Number
Warner Bros.	6.352519e+10	1.293792e+08	491
Universal Pictures	5.525919e+10	1.193503e+08	463
Paramount Pictures	4.880819e+10	1.235650e+08	395
Twentieth Century Fox Film Corporation	4.768775e+10	1.398468e+08	341
Walt Disney Pictures	4.083727e+10	2.778046e+08	147
Columbia Pictures	3.227974e+10	1.367785e+08	236
New Line Cinema	2.217339e+10	1.119868e+08	198
Amblin Entertainment	1.734372e+10	2.550547e+08	68
DreamWorks SKG	1.547575e+10	1.984071e+08	78
Dune Entertainment	1.500379e+10	2.419966e+08	62

- Warner Bros는 총 491편의 영화에서 635억 달러의 수입을 올린 제작사
- Univesal Pictures와 Paramaount Pictures는 각각 550억 달러, 480억 달러로 뒤를 잇는다

Production Countries

국가별 제작한 영화 개수 순위

	num_movies	country
0	21153	United States of America
1	4094	United Kingdom
2	3940	France
3	2254	Germany
4	2169	Italy
5	1765	Canada
6	1648	Japan
7	964	Spain
8	912	Russia
9	828	India

- 미국에서 제작된 영화가 가장 많음
- 아시아에서는 일본과 인도 순으로 많은 영화를 보유하고 있음

Original Language

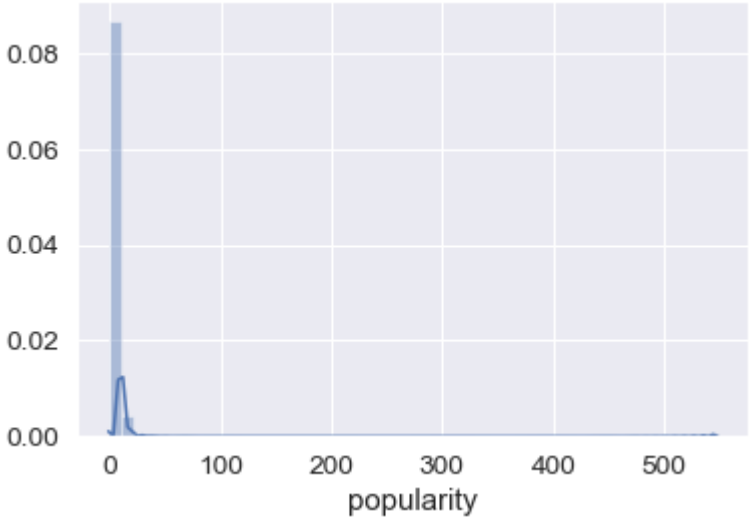
가장 많이 쓰인 언어 순위

	number	language
en	32269	en
fr	2438	fr
it	1529	it
ja	1350	ja
de	1080	de

- 영어가 압도적으로 많이 사용됨

Popularity

인기도 분포




```
count    45460.000000
mean         2.921478
std          6.005414
min          0.000000
25%          0.385948
50%          1.127685
75%          3.678902
max        547.488298
Name: popularity, dtype: float64
```

- 인기도는 평균 2.9점, 3사분위수가 3.67이지만, 최대값은 547점
- 극단적으로 skewed 되어 있음

Popularity

인기도가 높은 영화 순위

	title	popularity	year
30700	Minions	547.488298	2015
33356	Wonder Woman	294.337037	2017
42222	Beauty and the Beast	287.253654	2017
43644	Baby Driver	228.032744	2017
24455	Big Hero 6	213.849907	2014
26564	Deadpool	187.860492	2016
26566	Guardians of the Galaxy Vol. 2	185.330992	2017
14551	Avatar	185.070892	2009
24351	John Wick	183.870374	2014
23675	Gone Girl	154.801009	2014

- Minions, Wonder Woman, Beauty and the Beast가 각각 547점, 294점, 287점으로 높은 인기도를 보임

Vote Count

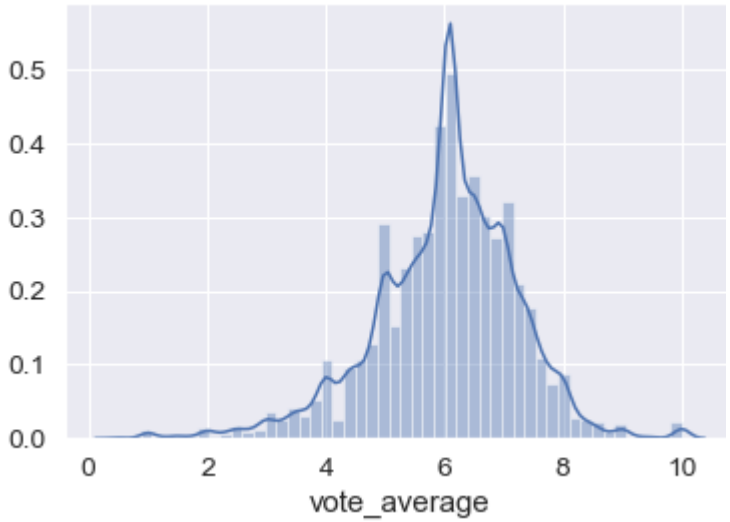
투표 수가 가장 많은 영화 순위

	title	vote_count	year
15480	Inception	14075.0	2010
12481	The Dark Knight	12269.0	2008
14551	Avatar	12114.0	2009
17818	The Avengers	12000.0	2012
26564	Deadpool	11444.0	2016
22879	Interstellar	11187.0	2014
20051	Django Unchained	10297.0	2012
23753	Guardians of the Galaxy	10014.0	2014
2843	Fight Club	9678.0	1999
18244	The Hunger Games	9634.0	2012

- Inception과 The Dark Knight가 가장 많이 투표된 영화
- 두 영화 모두 크리스토퍼 놀란 作

Vote Average

평균 평점 분포




```
count    42462.000000
mean         6.014877
std          1.256208
min          0.500000
25%          5.300000
50%          6.100000
75%          6.900000
max         10.000000
Name: vote_average, dtype: float64
```

- 평점 평균은 10점 만점에 6점

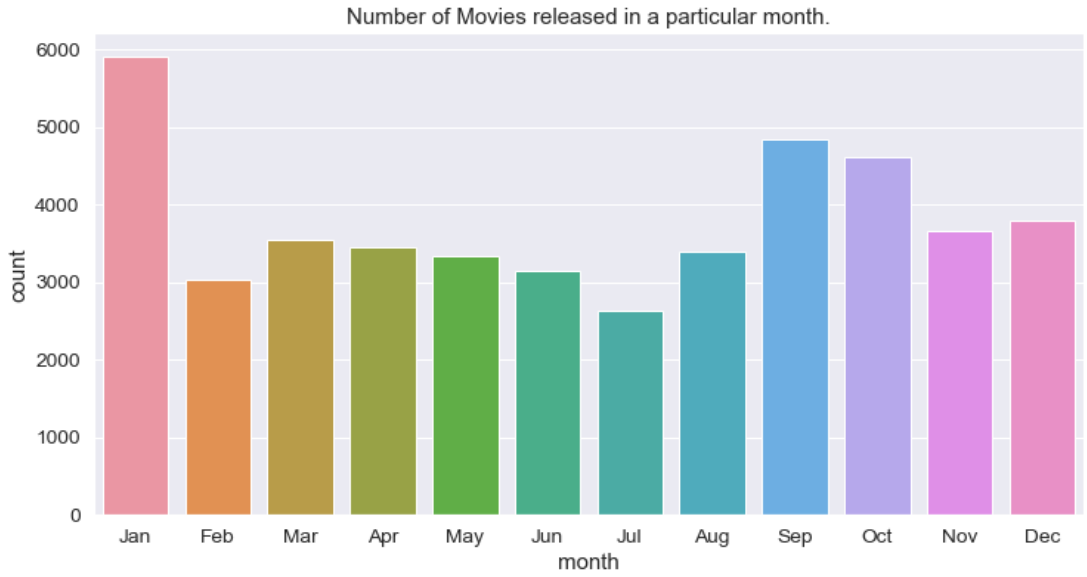
Vote Average

2000표 이상 투표된 영화의 평균 평점 순위

	title	vote_average	vote_count	year
314	The Shawshank Redemption	8.5	8358.0	1994
834	The Godfather	8.5	6024.0	1972
2211	Life Is Beautiful	8.3	3643.0	1997
5481	Spirited Away	8.3	3968.0	2001
1152	One Flew Over the Cuckoo's Nest	8.3	3001.0	1975
1176	Psycho	8.3	2405.0	1960
2843	Fight Club	8.3	9678.0	1999
1178	The Godfather: Part II	8.3	3418.0	1974
12481	The Dark Knight	8.3	12269.0	2008
292	Pulp Fiction	8.3	8670.0	1994

Movie Release Dates

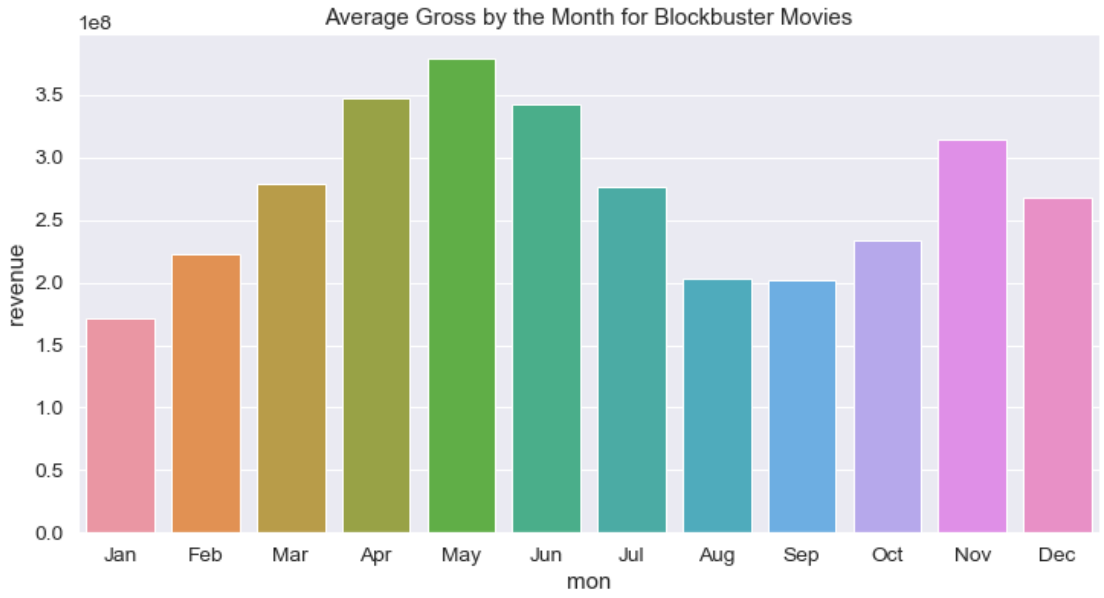
월별 개봉한 영화 개수



- 1월에 가장 많은 영화가 개봉함
- 할리우드에서는 인디 영화 12개가 개봉하는 dump month이기도 함

Movie Release Dates

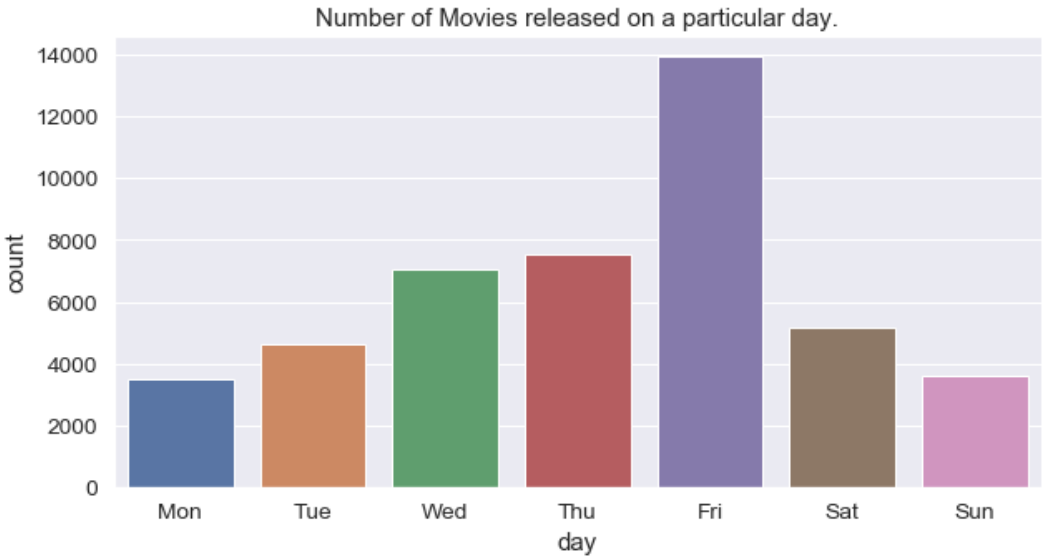
월별 평균 수익



- 4~6월에 개봉한 영화가 수익이 높았음
- 방학이나 휴가의 영향이 있을 것으로 예상
- 블록버스터 영화 출시와 관련이 있음

Movie Release Dates

요일별 개봉한 영화 개수



- 주말의 시작인 금요일에 가장 많이 출시됨
- 같은 이유로 일요일과 월요일에 가장 적게 출시

Movie Release Dates

개봉한지 오래된 순위

	title	year
34940	Passage of Venus	1874
34937	Sallie Gardner at a Gallop	1878
41602	Buffalo Running	1883
34933	Man Walking Around a Corner	1887
34934	Accordion Player	1888
34938	Traffic Crossing Leeds Bridge	1888
34936	Monkeyshines, No. 2	1890
34939	London's Trafalgar Square	1890
34935	Monkeyshines, No. 1	1890
41194	Mosquinha	1890

- 데이터 세트에 있는 영화 중 가장 오래된 영화는 Passage of Venus
- 1874년 개봉, 프랑스 천문학자인 Pierre Janssen이 촬영
- 태양을 가로 지르는 금성을 촬영한 것

Movie Status

출시 여부에 대한 분포

```
Released            45014
Rumored               230
Post Production        98
In Production          20
Planned                15
Canceled                2
Name: status, dtype: int64
```

- MovieLens에는 아직 계획 중이거나,
  제작 및 출시 전 단계의 영화에 대한 데이터도 있음

Spoken Languages

사용된 언어의 개수에 대한 분포

```
1     33736
2      5371
0      3835
3      1705
4       550
5       178
6        62
7        14
8         6
9         5
19        1
13        1
12        1
10        1
Name: spoken_languages, dtype: int64
```

- 대부분의 영화가 하나의 언어만 사용
- 19개 언어가 최다 사용 언어 수

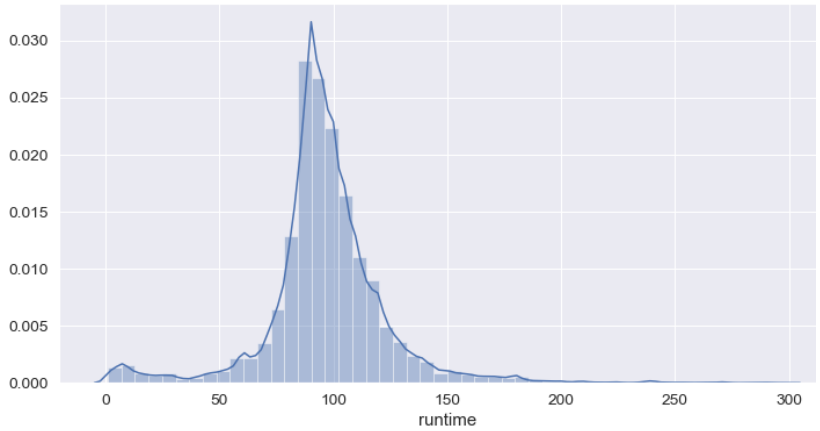
Spoken Languages

사용한 언어 개수가 많은 영화 순위

	title	year	spoken_languages
22235	Visions of Europe	2004	19
35288	The Testaments	2000	13
14093	To Each His Own Cinema	2007	12
8789	The Adventures of Picasso	1978	10

- 가장 많은 언어를 사용한 Visions of Europe은 25명의 유럽 감독이 제작한 25편의 단편 영화 모음

Runtime




```
count    45203.000000
mean        94.128199
std         38.407810
min          0.000000
25%         85.000000
50%         95.000000
75%        107.000000
max       1256.000000
Name: runtime, dtype: float64
```

- 평균 런타임은 약 1시간 30분 정도
- 가장 긴 영화는 1256분(대략 21시간)
- 오른쪽 그래프는 런타임이 300분 미만인 영화들의 분포를 나타냄

Runtime

런타임이 짧은 순서

	runtime	title	year
35146	1.0	La Vague	1891
36577	1.0	Champs de Mars	1900
36576	1.0	Palace of Electricity	1900
44965	1.0	Luis Martinetti, Contortionist	1894
42357	1.0	A Gathering of Cats	2007
36575	1.0	Eiffel Tower from Trocadero Palace	1900
36574	1.0	Panorama of Eiffel Tower	1900
44714	1.0	The Infernal Caldron	1903
19244	1.0	The Kiss	1896
44646	1.0	The Vanishing Lady	1896

런타임이 긴 순서

	runtime	title	year
24178	1256.0	Centennial	1978
40938	1140.0	Baseball	1994
19965	1140.0	Jazz	2001
13767	931.0	Berlin Alexanderplatz	1980
13953	925.0	Heimat: A Chronicle of Germany	1984
27855	900.0	The Story of Film: An Odyssey	2011
26682	877.0	Taken	2002
19158	874.0	The War	2007
34667	840.0	The Roosevelts: An Intimate History	2014
34732	840.0	Seventeen Moments in Spring	1973

- 1890년대 후반~1900년대 초반에 1분 길이의 영화가 많음
- 런타임이 긴 순으로 랭크된 것들은
  대부분 TV 미니 시리즈인 것으로 사료됨

Budget

가장 많은 예산을 쓴 영화 순위

	title	budget	revenue	year
17124	Pirates of the Caribbean: On Stranger Tides	380000000.0	1.045714e+09	2011
11827	Pirates of the Caribbean: At World's End	300000000.0	9.610000e+08	2007
26558	Avengers: Age of Ultron	280000000.0	1.405404e+09	2015
11067	Superman Returns	270000000.0	3.910812e+08	2006
44842	Transformers: The Last Knight	260000000.0	6.049421e+08	2017
16130	Tangled	260000000.0	5.917949e+08	2010
18685	John Carter	260000000.0	2.841391e+08	2012
11780	Spider-Man 3	258000000.0	8.908716e+08	2007
21175	The Lone Ranger	255000000.0	8.928991e+07	2013
22059	The Hobbit: The Desolation of Smaug	250000000.0	9.584000e+08	2013

- 두 개의 Pirates of the Caribbean 시리즈가 3억 달러가 넘는 예산으로 1위
- Lone Ranger를 제외하고는 모두 예산보다 높은 수익을 거둠

Revenue

고수익을 거둔 영화 순위

	poster_path	title	budget	revenue	year
14551		Avatar	237000000.0	2.787965e+09	2009
26555		Star Wars: The Force Awakens	245000000.0	2.068224e+09	2015
1639		Titanic	200000000.0	1.845034e+09	1997

- 각각 1위와 3위인 Avatar와 Titanic은 James Cameron 作

Genres

	genre	movies
0	Drama	20265
1	Comedy	13182
2	Thriller	7624
3	Romance	6735
4	Action	6596
5	Horror	4673
6	Crime	4307
7	Documentary	3932
8	Adventure	3496
9	Science Fiction	3049

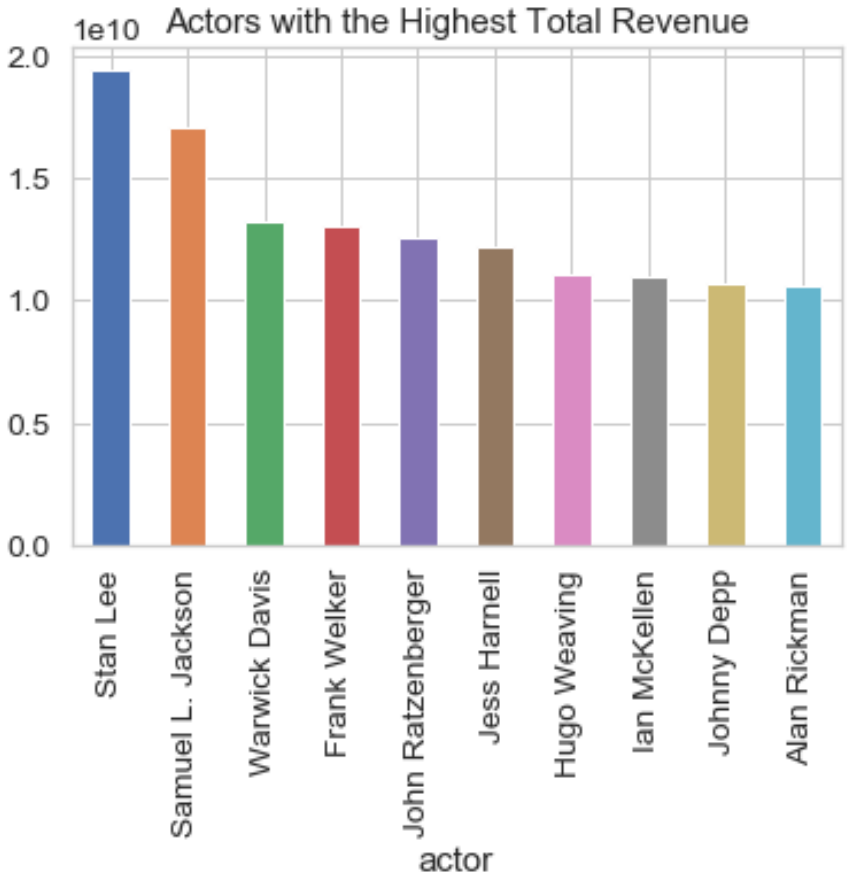
- 총 32개의 장르
- 데이터 세트의 약 절반 가량의 영화가 드라마 장르

credits

cast	출연진 이름과 역할에 대한 문자열 딕셔너리
crew	제작진 이름과 수행한 역할에 대한 문자열 딕셔너리
id	TMDb ID

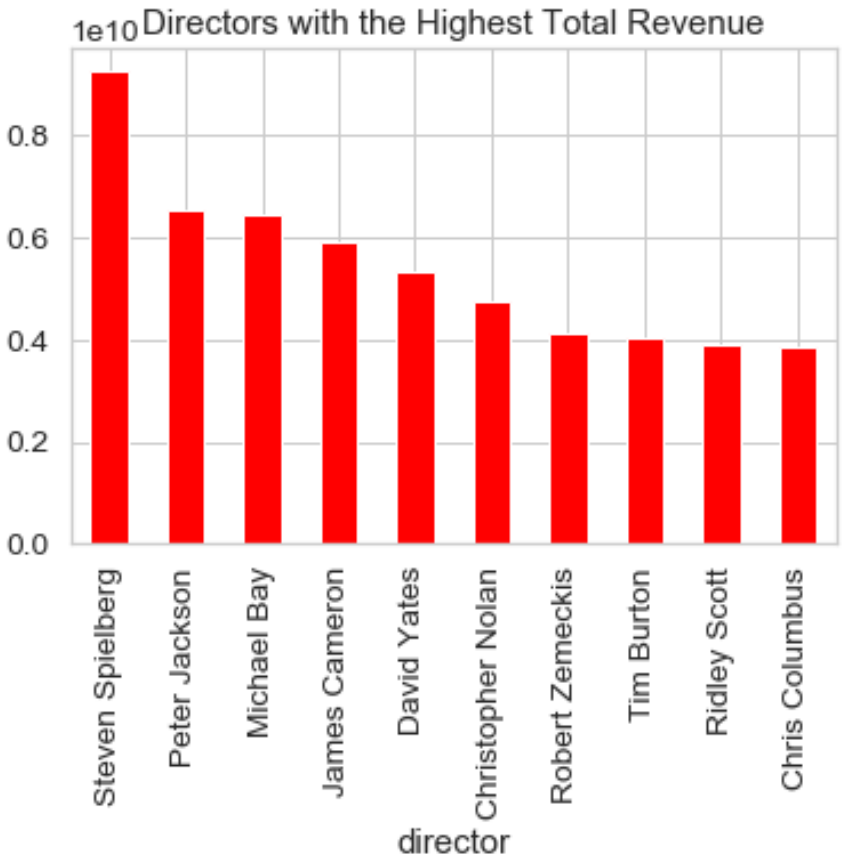
Cast

총 수입이 높은 배우 순서



Crew

총 수입이 높은 감독 순서



Recommender System

사용 알고리즘

- Demographic Filtering
- Content based Recommender

  1) Movie description based

  2) Meta data based
- Popularity and Ratings based Recommender
- Collaborative filtering
- Hybrid Recommender

Demographic Filtering

- 일반적인 추천 정보를 제공

- 점수를 매기거나 영화를 평가할 척도가 필요함
- 평균 평점을 점수로 사용하는 것은 적합하지 않음
  (3개 투표로 평균 8점인 영화 vs 50개의 투표로 평균 7점인 영화)
- 따라서 IMDB의 가중치 사용

Demographic Filtering

- IMDB의 가중치 적용 공식 사용하여 차트 구성
- 가중치(WR) = $\left(\frac{v}{v+m} \cdot R\right) + \left(\frac{m}{v+m} \cdot C\right)$
- v : 총 투표 수
- m : 차트에 나열되기 위해 필요한 최소 투표 수
- R : 영화의 평균 평점
- C : 전체 데이터 세트에서 평균 평점

Demographic Filtering

평점이 높은 영화 순위

```
qualified.head(15)
```

	title
15480	Inception
12481	The Dark Knight
22879	Interstellar
2843	Fight Club
4863	The Lord of the Rings: The Fellowship of the Ring
292	Pulp Fiction
314	The Shawshank Redemption
7000	The Lord of the Rings: The Return of the King
351	Forrest Gump
5814	The Lord of the Rings: The Two Towers
256	Star Wars
1225	Back to the Future
834	The Godfather
1154	The Empire Strikes Back
46	Se7en

장르별 평점 순위 (로맨스)

```
build_chart('Romance').head(15)
```

	title
10309	Dilwale Dulhania Le Jayenge
351	Forrest Gump
876	Vertigo
40251	Your Name.
883	Some Like It Hot
1132	Cinema Paradiso
19901	Paperman
37863	Sing Street
882	The Apartment
38718	The Handmaiden
3189	City Lights
24886	The Way He Looks
45437	In a Heartbeat
1639	Titanic
19731	Silver Linings Playbook

Content Based Recommender

- 영화 간의 유사성을 계산하여 사용자가 좋아하는 특정 영화와 가장 유사한 영화를 제안하는 엔진

- Movie Description Based
- Metadata Based

Content Based Recommender

- Cosine Similarity를 사용하여 두 영화 사이의 유사성 측정
- $cosine(x, y) = \frac{x \cdot y^T}{||x|| \cdot ||y||}$
- sklearn의 linear_kernel사용하여 계산 가능

Movie Description Based

- Overviews, taglines 사용

```
get_recommendations('The Dark Knight').head(10)
```

```
7931                       The Dark Knight Rises
132                              Batman Forever
1113                             Batman Returns
8227    Batman: The Dark Knight Returns, Part 2
7565                 Batman: Under the Red Hood
524                                      Batman
7901                          Batman: Year One
2579               Batman: Mask of the Phantasm
2696                                        JFK
8165    Batman: The Dark Knight Returns, Part 1
Name: title, dtype: object
```

- The Dark Knight를 예시로 하면,
  구현한 시스템은 이 영화를 배트맨 영화로 식별하였음

Metadata Based

- Cast, crew, keywords, genre 사용

- Cast에서는 목록의 상위 3명의 배우만 선택
- Crew에서는 감독만 추출
- Keywords – 한번만 발생하는 키워드는 제외
  동일한 단어를 한가지로 인식하도록 변환
  (ex. Dogs와 dog를 같은 단어로 간주)

- 감독 및 배우, 장르가 Overviews와 taglines보다 영화의 평판이나 인기를 좌우할 것으로 사료됨
- 따라서 앞선 내용 기반 시스템보다 유용할 것으로 예상

Metadata Based

```
get_recommendations('The Dark Knight').head(10)
```

```
8031          The Dark Knight Rises
6218                Batman Begins
6623                The Prestige
2085                   Following
7648                   Inception
4145                    Insomnia
3381                    Memento
8613                Interstellar
7659    Batman: Under the Red Hood
1134                Batman Returns
Name: title, dtype: object
```

- The Dark Knight를 입력했을 때, Batman Begins, The Prestige, Dark Knight Rises 등 다른 Christopher Nolan의 영화를 추천해줌
- 그러나 평점이나 인기도에는 관계없이 추천한다는 단점

Popularity and Ratings Based Recommender

- 컨텐츠 기반 추천 시스템의 단점 보완
- 평점이 낮은 영화는 제거하고, 대중적인 영화를 반환하는 매커니즘을 추가

Popularity and Ratings Based Recommender

```
improved_recommendations('The Dark Knight')
```

	title	vote_count	vote_average	year	wr
7648	Inception	14075	8	2010	7.917588
8613	Interstellar	11187	8	2014	7.897107
6623	The Prestige	4510	8	2006	7.758148
3381	Memento	4168	8	2000	7.740175
8031	The Dark Knight Rises	9263	7	2012	6.921448
6218	Batman Begins	7511	7	2005	6.904127
1134	Batman Returns	1706	6	1992	5.846862
132	Batman Forever	1529	5	1995	5.054144
9024	Batman v Superman: Dawn of Justice	7189	5	2016	5.013943
1260	Batman & Robin	1447	4	1997	4.287233

- 사용자의 개인적인 취향은 반영되지 않는 단점

Collaborative Filtering

- 나와 비슷한 취향을 가진 사용자들이
  어떤 영화를 많이 봤는지에 따라 추천
- Surprise library 사용
  (RMSE를 최소화 하기 위하여 SVD 알고리즘 사용)

Collaborative Filtering

```
Evaluating RMSE, MAE of algorithm SVD.

------------
Fold 1
RMSE: 0.9020
MAE:  0.6944
------------
Fold 2
RMSE: 0.8883
MAE:  0.6838
------------
Fold 3
RMSE: 0.8906
MAE:  0.6887
------------
Fold 4
RMSE: 0.9006
MAE:  0.6941
------------
Fold 5
RMSE: 0.8967
MAE:  0.6889
------------
------------
Mean RMSE: 0.8956
Mean MAE : 0.6900
------------
------------
```

```
CaseInsensitiveDefaultDict(list,
                           {'rmse': [0.9019645118119045,
                             0.8882964405241228,
                             0.8905754652687611,
                             0.9006036838183014,
                             0.896714544718544],
                            'mae': [0.6943760609152614,
                             0.6837909215632183,
                             0.6887055741532625,
                             0.694132027614935,
                             0.6888988930689208]})
```

- 대략 0.89~0.9 사이의 RMSE

Collaborative Filtering

	userId	movieId	rating	timestamp
0	1	31	2.5	1260759144
1	1	1029	3.0	1260759179
2	1	1061	3.0	1260759182
3	1	1129	2.0	1260759185
4	1	1172	4.0	1260759205
5	1	1263	2.0	1260759151
6	1	1287	2.0	1260759187
7	1	1293	2.0	1260759148
8	1	1339	3.5	1260759125
9	1	1343	2.0	1260759131
10	1	1371	2.5	1260759135

```
svd.predict(1, 302, 3)
```

```
Prediction(uid=1, iid=302, r_ui=3, est=2.654124992619001, details={'was_impossible': False})
```

- 1번 user에 대하여 ID가 302인 영화의 경우, 예상 예측치는 2.654

Hybrid Recommender

- 콘텐츠 기반과 협업 필터 기반 시스템 결합

- 작동 방법

  Input : 사용자 ID 및 영화 제목

  Output : 사용자가 줄 것으로 예상되는 평점 기준으로
  비슷한 영화들을 정렬

Hybrid Recommender

```
hybrid(1, 'Avatar')
```

	title
1011	The Terminator
974	Aliens
522	Terminator 2: Judgment Day
2014	Fantastic Planet
8401	Star Trek Into Darkness
8658	X-Men: Days of Future Past
1621	Darby O'Gill and the Little People
1668	Return from Witch Mountain
922	The Abyss
3060	Sinbad and the Eye of the Tiger

```
hybrid(500, 'Avatar')
```

	title
974	Aliens
2014	Fantastic Planet
4966	Hercules in New York
8401	Star Trek Into Darkness
1621	Darby O'Gill and the Little People
1668	Return from Witch Mountain
3060	Sinbad and the Eye of the Tiger
922	The Abyss
4347	Piranha Part Two: The Spawning
1376	Titanic

- 같은 영화도 사용자마다 다른 추천 목록을 얻음
- 보다 개인 맞춤화된 시스템 구현

요약

- Demographic Filtering

  투표 수와 평균 점수를 사용하여 특정 장르의 인기 영화 차트 작성

- Content Based Recommender

  1) Movie description based - overviews, taglines 이용

  2) Meta data based - cast, crew, genre, keywords 이용

- Popularity and Ratings Based Recommender

  인기도와 평점 고려하여 보다 대중적인 영화 추천

- Collaborative Filtering

  Surprise library 사용하여 사용자의 특정 영화에 대한 예상 평점 제공

- Hybrid Recommender

  콘텐츠 및 협업 필터링 결합

  예상 평점을 기반으로 추천 목록을 제공하는 엔진을 만듦