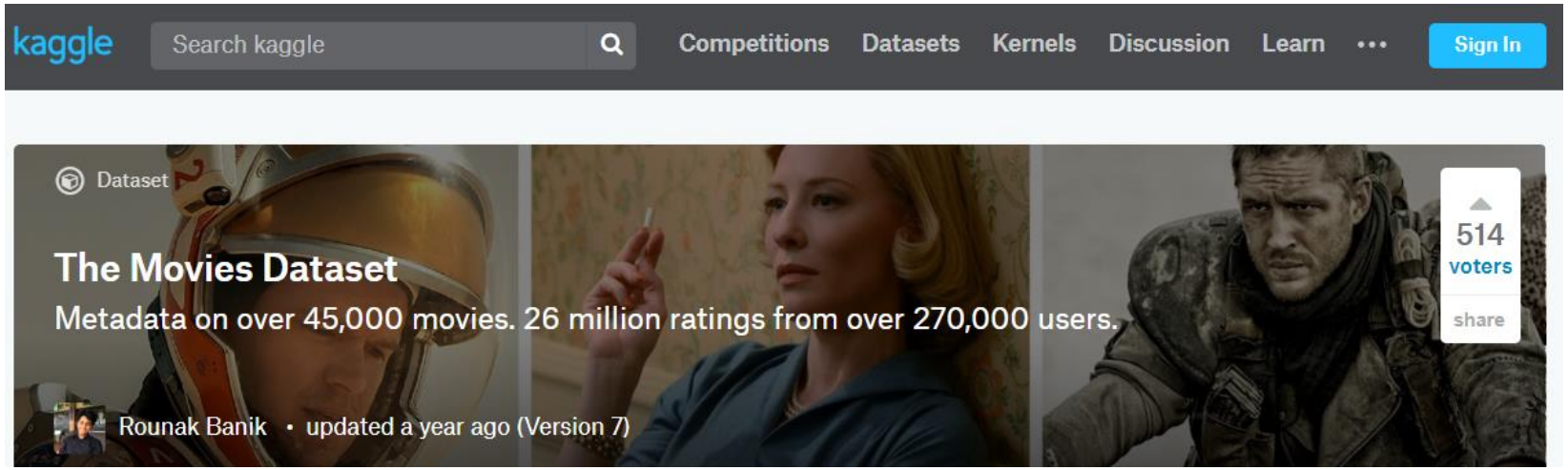
키워드에 따른 영화의 수익 예측

박예빈 박해영 유정아 정상아

1. Raw Data



45000개 이상의 영화를 분석한 Dataset

원출처 : TMDB(The Movie DataBase 사이트)

1. Raw Data

	title	genres
2	Toy Story	[{'id': 16, 'name': 'Animation'}, {'id': 35, 'name': 'Comedy'}, {'id': 10751, 'name': 'Family'}]
3	Jumanji	[{'id': 12, 'name': 'Adventure'}, {'id': 14, 'name': 'Fantasy'}, {'id': 10751, 'name': 'Family'}]
4	Grumpier Old Men	[{'id': 10749, 'name': 'Romance'}, {'id': 35, 'name': 'Comedy'}]
5	Waiting to Exhale	[{'id': 35, 'name': 'Comedy'}, {'id': 18, 'name': 'Drama'}, {'id': 10749, 'name': 'Romance'}]

	keywords
2	[{'id': 931, 'name': 'jealousy'}, {'id': 4290, 'name': 'toy'}, {'id': 5202, 'name': 'boy'}, {'id': 6054, 'name': 'friendship'},
3	[{'id': 10090, 'name': 'board game'}, {'id': 10941, 'name': 'disappearance'}, {'id': 15101, 'name': "based on childr
4	[{'id': 1495, 'name': 'fishing'}, {'id': 12392, 'name': 'best friend'}, {'id': 179431, 'name': 'duringcreditsstinger'}, {'id
5	[{'id': 818, 'name': 'based on novel'}, {'id': 10131, 'name': 'interracial relationship'}, {'id': 14768, 'name': 'single n
6	[{'id': 1009, 'name': 'baby'}, {'id': 1599, 'name': 'midlife crisis'}, {'id': 2246, 'name': 'confidence'}, {'id': 4995, 'nam

출연진, 장르, 줄거리 키워드, 예산, 수익

개봉연도, 제작 회사, 국가, TMDB 투표 수 및 투표 평균

2. 목표

프로젝트 목표

영화 평가 사이트의 투표수 & 평점을 이용

특정 장르에서 어떤 키워드를 가진 영화가 얼마의 수익을 가질 것인지 예측

1. Raw Data

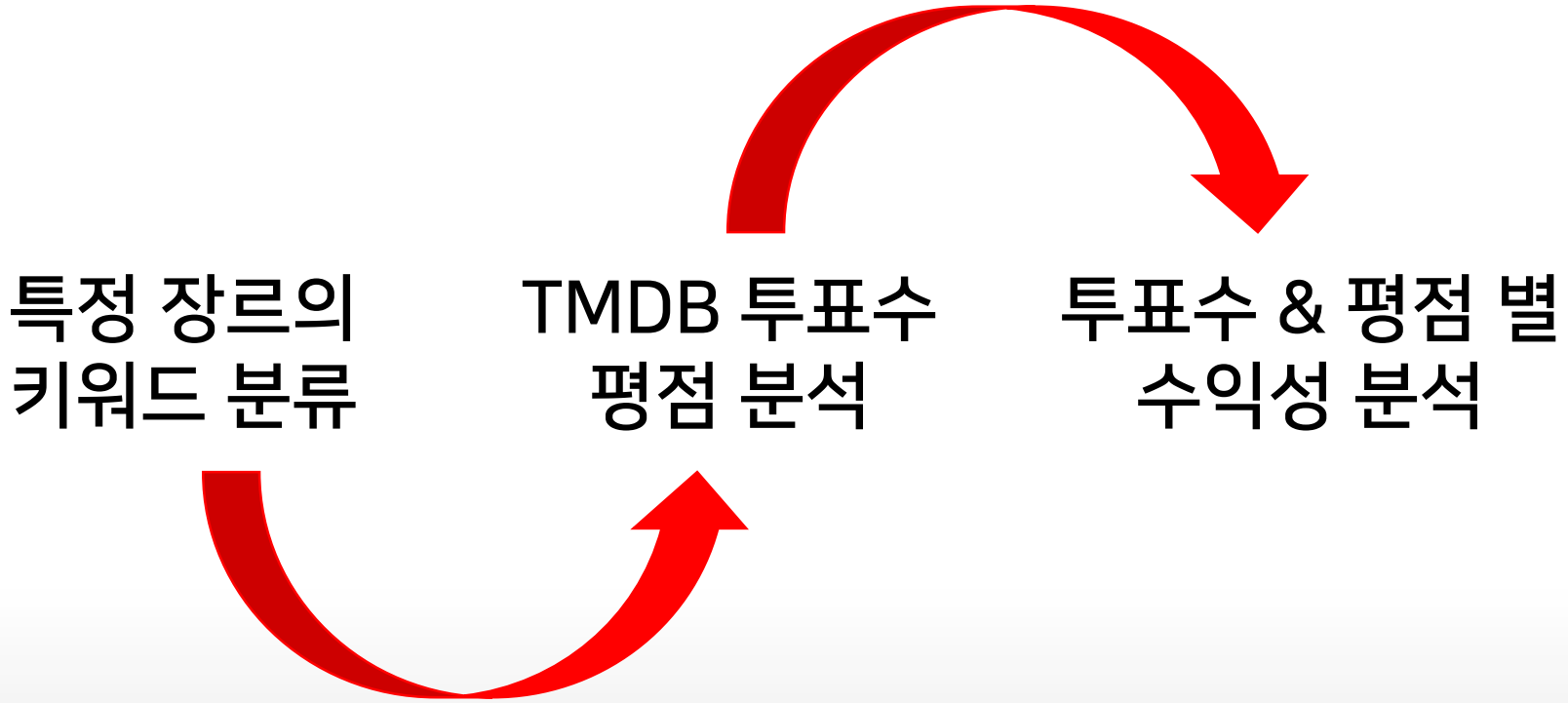
	title	budget	revenue	vote_averag	vote_count
2	Toy Story	30000000	373554033	7.7	5415
3	Jumanji	65000000	262797249	6.9	2413
4	Grumpier Old Men	0	0	6.5	92
5	Waiting to Exhale	16000000	81452156	6.1	34
6	Father of the Bride Part II	0	76578911	5.7	173
7	Heat	60000000	187436818	7.7	1886
8	Sabrina	58000000	0	6.2	141
9	Tom and Huck	0	0	5.4	45
10	Sudden Death	35000000	64350171	5.5	174
11	GoldenEye	58000000	352194034	6.6	1194

추출할 데이터 종류

- 장르와 키워드
- 투표수 & 평점
- 예산과 수익

3. 데이터 처리

분석 과정



3. 데이터 처리

분석 과정

특정 장르의
키워드

수익성 예측

3. 데이터 처리

3) 데이터 추출 & Remove Outliers

	keywords
2	[{'id': 931, 'name': 'jealousy'}, {'id': 4290, 'name': 'toy'}, {'id': 5202, 'name': 'boy'}, {'id': 6054, 'name': 'friendship'},
3	[{'id': 10090, 'name': 'board game'}, {'id': 10941, 'name': 'disappearance'}, {'id': 15101, 'name': "based on childr
4	[{'id': 1495, 'name': 'fishing'}, {'id': 12392, 'name': 'best friend'}, {'id': 179431, 'name': 'duringcreditsstinger'}, {'id
5	[{'id': 818, 'name': 'based on novel'}, {'id': 10131, 'name': 'interracial relationship'}, {'id': 14768, 'name': 'single n
6	[{'id': 1009, 'name': 'baby'}, {'id': 1599, 'name': 'midlife crisis'}, {'id': 2246, 'name': 'confidence'}, {'id': 4995, 'nam

Text(Keyword) 추출

3. 데이터 처리

1) 군집화 – LDA 모델링 (Latent Dirichlet Allocation)

: 문서에 포함된 단어를 빈도수에 따라 서로 연관 짓고 분류

LDA Topic 1

LDA Topic 2

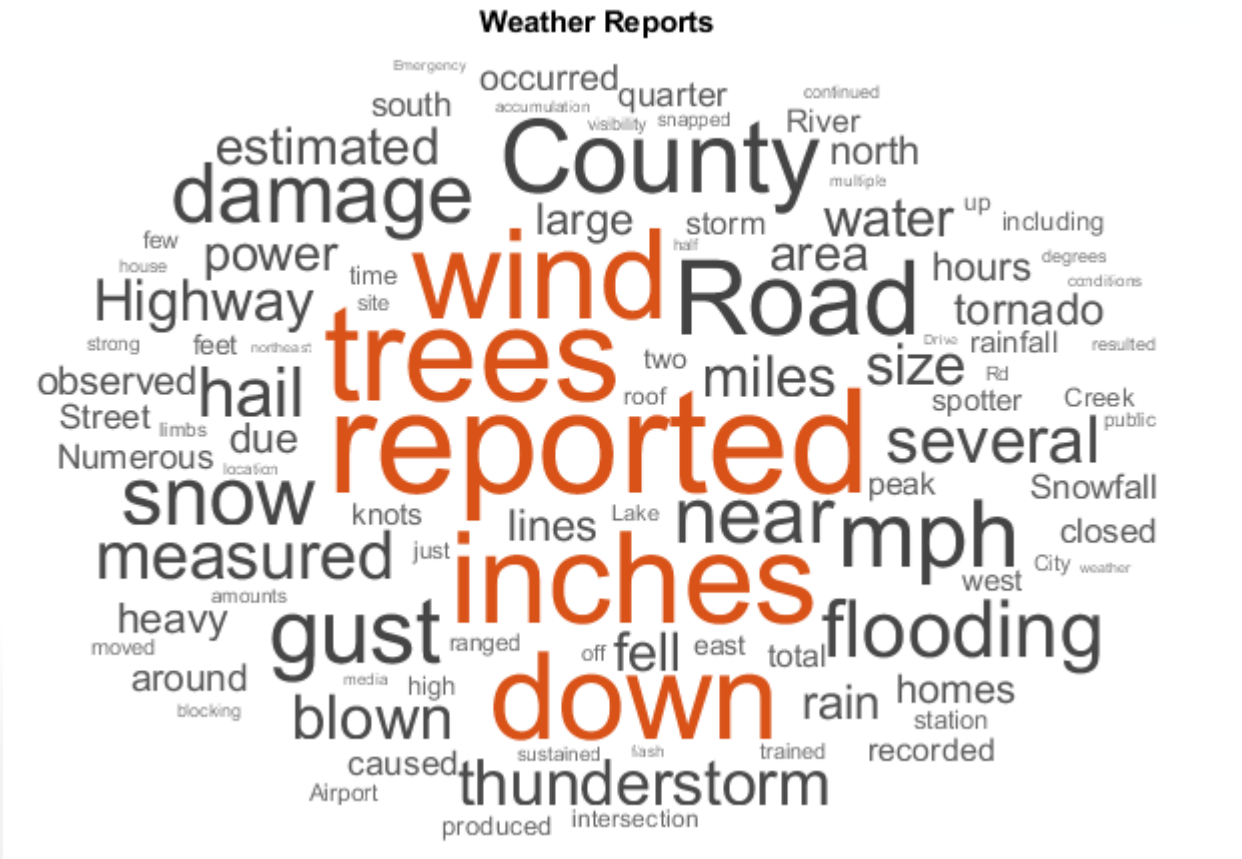
LDA Topic 3

LDA Topic 4

3. 데이터 처리

1) 키워드 분석 – Visualize Text Data Using Word Clouds

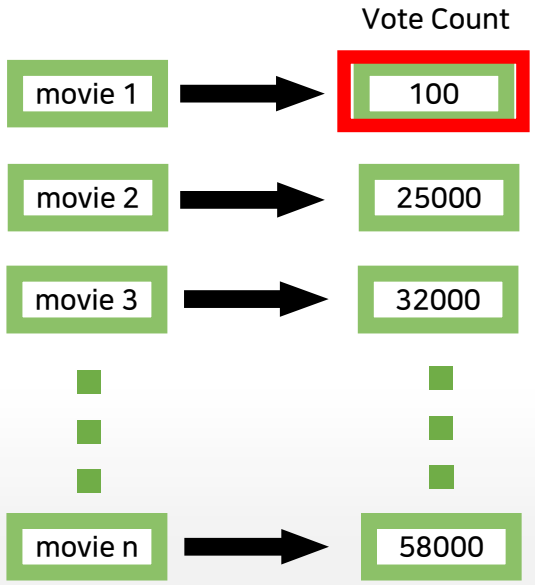
: Word clouds 를 사용한 텍스트 데이터 시각화



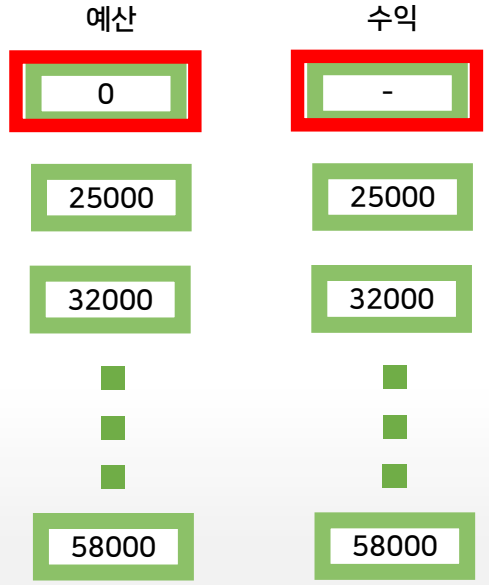
3. 데이터 처리

3) 데이터 추출 & Remove Outliers

평점의 신뢰성 구축을 위해
극단치(Outlier) 제거

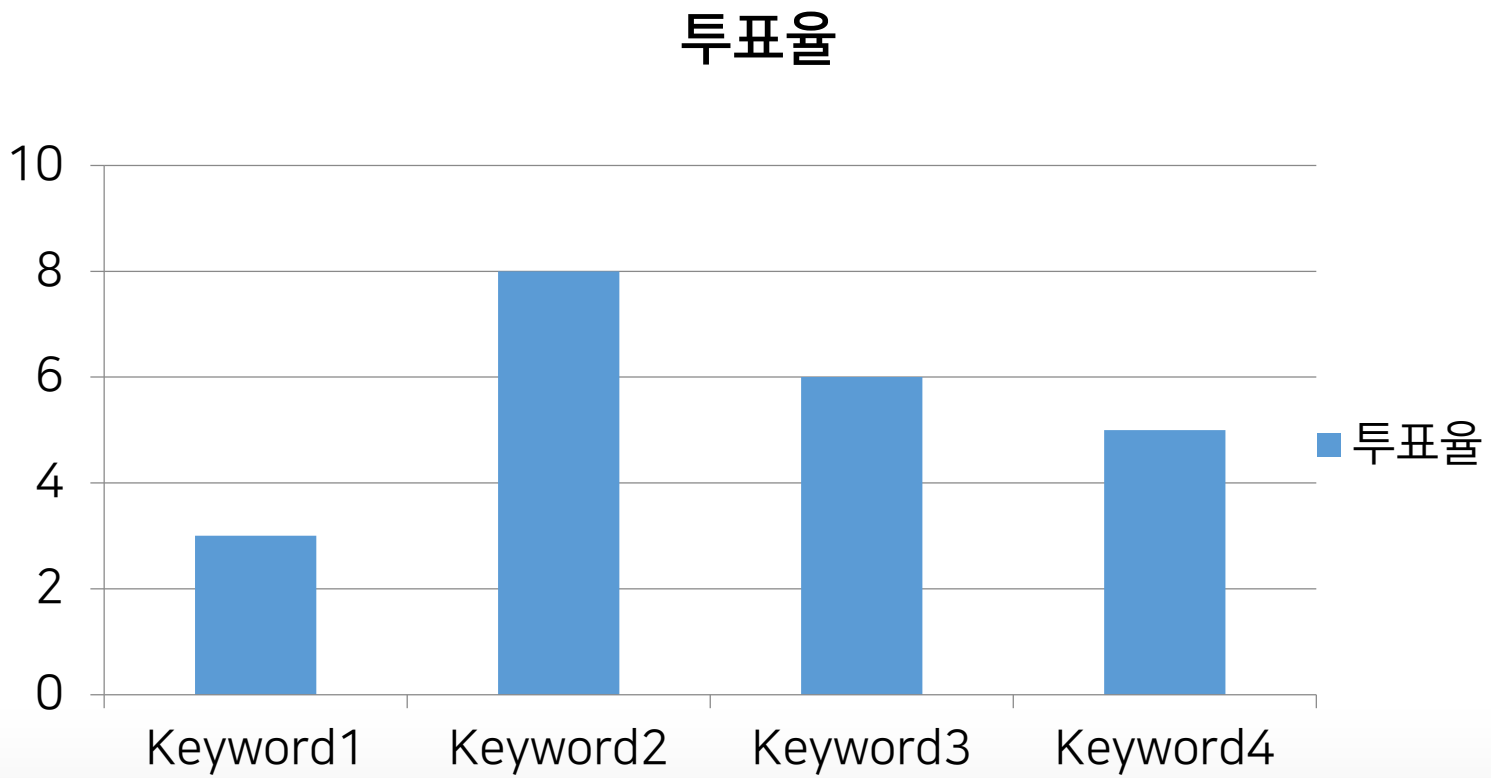


측정할 수 없는 값은 배제
(수익과 예산이 0인 항목)



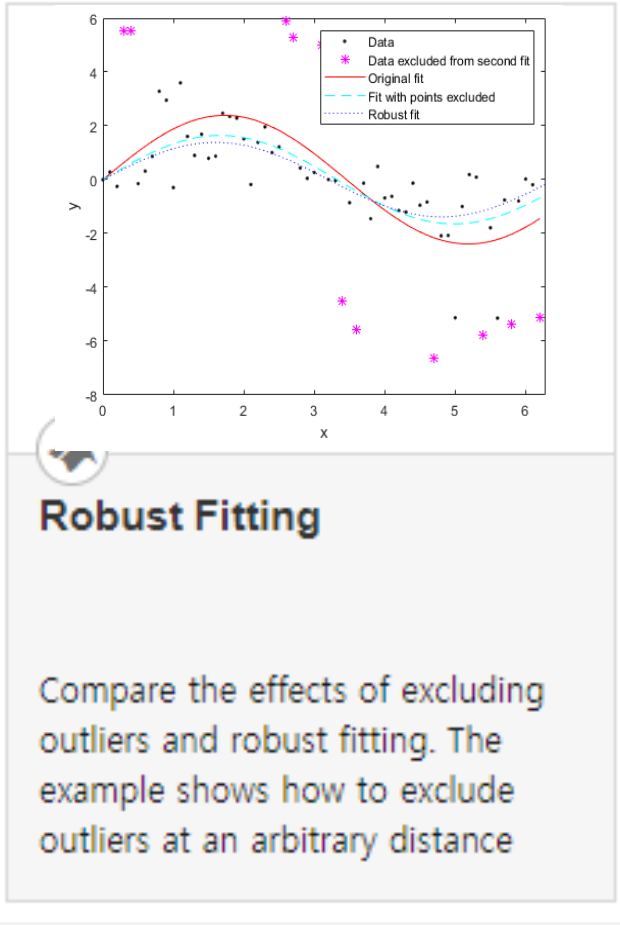
3. 데이터 처리

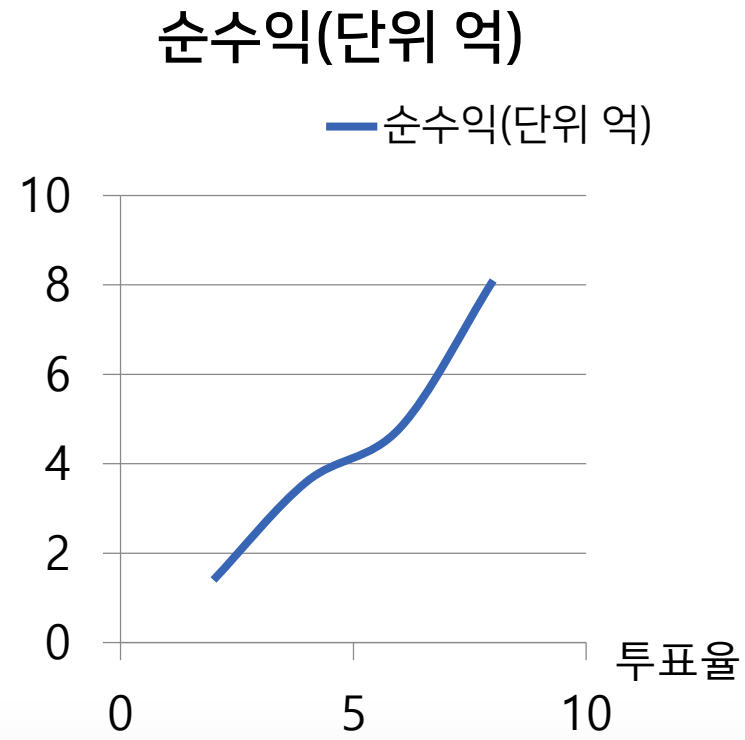
2) 각 군집(Topic)에 따른 투표율 & 평점 그래프



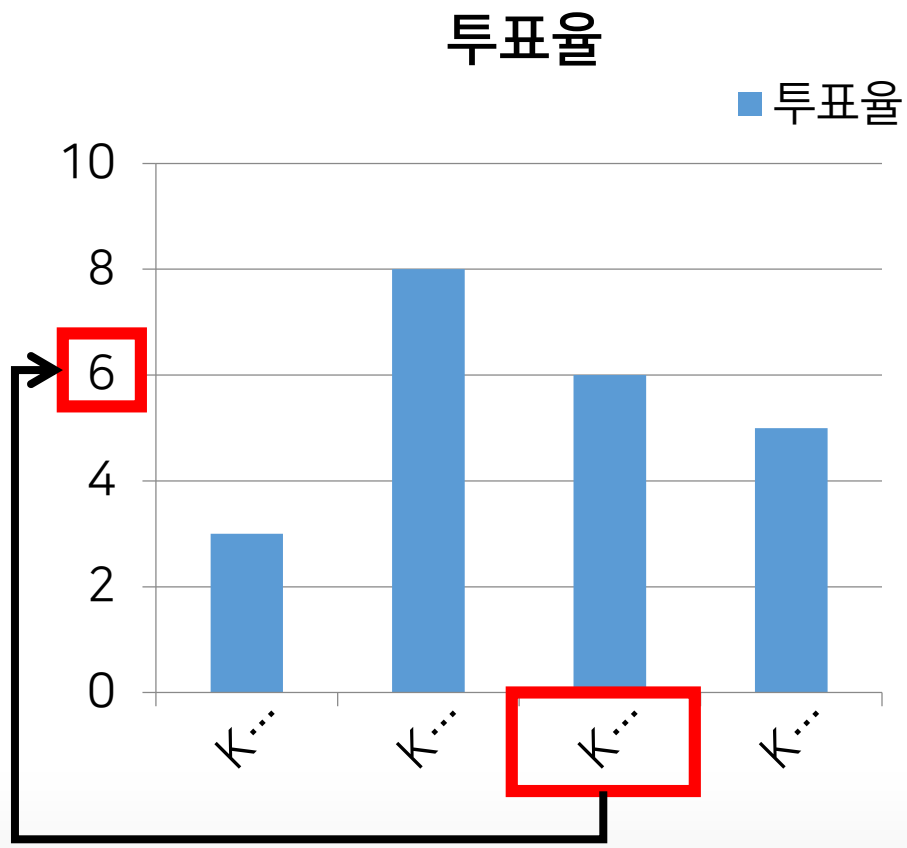
3. 데이터 처리

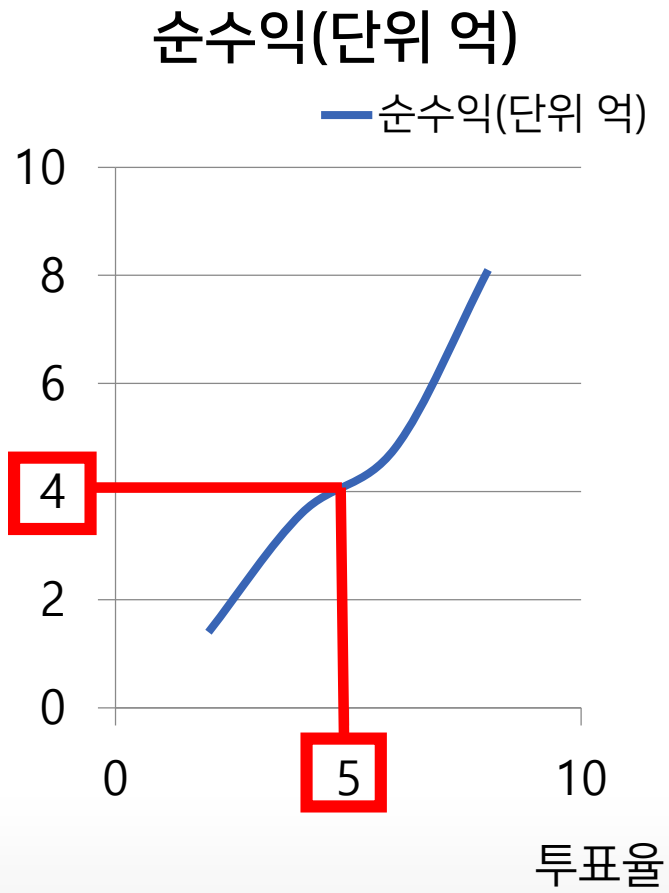
4) Curve fitting / 투표율 & 평점에 따른 순수익 그래프





4. 예측가능한 결과





5. 기대효과

기대효과

제작자 또는 배우의 입장에서 특정 장르의 영화를 제작 & 출연하고자 할 때,

성공하기 위해 어떤 **키워드**를 가진 영화를 제작 & 출연해야 하는가

감사합니다.