Final Report (Implementation)

2023713107 장병우

1. What: Data Abstraction

- Brief description
  MovieLens 1M Dataset, 2000년 ~ 2003년까지 집계된 영화 평점 데이터이며 table 형태로 제공되어 있음.
  user 정보, item 정보, rating 정보 3가지 파일로 구성되어 있으며,
  각각 6,040개, 3,706개, 1,000,209개의 row 존재함

- item (영화) 정보가 너무 단순해서, IMDB와의 연계를 통해 추가적으로 감독, 배우 정보를 획득

- Abstraction

(IMDB 연동 전)

Number of users	6,040			
Number of items	3,706			
Number of ratings	1,000,209			
Sparsity	95.5316%			
User contents	**Gender**	**Age**	**Occupation**	**Zip code**
User contents dtype	attribute	attribute	attribute	attribute
User contents range	M, F	1, 18, 25, 35, 45, 50, 56	0~20	02460, 55117, ..., 70072
User contents mean	Male, Female	1: 1~17세 18: 18~24세	숫자=직업 종류	3,402개 우편번호
Item contents	**Genre**		**Title**	
Item contents dtype	attribute		attribute	
Item contents range	Action, Adeventure, ... , War		Toy Story, ... , Jumanji	
Item contents mean	어떤 장르에 해당하는가		3,706개의 영화 제목	
Rating contents	**Rating**			
Rating contents dtype	link			
Rating contents range	1, 2, 3, 4, 5			
Rating contents mean	유저가 영화에 매긴 1점과 5점 사이의 평점			

(IMDB 연동 후)

Number of users Number of items Number of ratings Sparsity	6,040 3,706 1,000,209 95.5316%			
User contents	**Gender**	**Age**	**Occupation**	**Zip code**
User contents dtype	attribute	attribute	attribute	attribute
User contents range	M, F	1, 18, 25, 35, 45, 50, 56	0~20	02460, 55117, ..., 70072
User contents mean	Male, Female	1: 1~17세 18: 18~24세	숫자=직업 종류	3,402개 우편번호
Item contents	**Genre**	**Director**	**Actor**	
Item contents dtype	attribute	attribute	attribute	
Item contents range	Action, Adeventure, ... , War	akira kurosawa, ... , zalman king	al pacino, ... , zahra naderi	
Item contents mean	어떤 장르에 해당하는가	어떤 감독이 연출했는가	어떤 배우가 연기했는가	
Rating contents	**Rating**			
Rating contents dtype	link			
Rating contents range	1, 2, 3, 4, 5			
Rating contents mean	유저가 영화에 매긴 1점과 5점 사이의 평점			

2. Who: Creating Personas

	Emily	John
Occupation	Data Scientist	Marketing Manager
Age	29	35
Location	Boston	San Francisco
Needs	Emily는 영화 추천시스템을 연구하는데, 인구통계적 데이터와 유저의 평점 간의 상세한 correlation을 확인할 수 있는 시스템이 필요함	John은 다양한 연령대의 인기 장르를 기반으로 한 마케팅 캠페인을 목표로 하고 있는데, 이때 장르의 인기와 사용자 참여에 대한 insight가 필요함.
Pain Points	분류되지 않은 대량의 데이터 때문에, 추세를 파악하기 어려움	현재 시장 트렌드와 부합하는 실시간 데이터를 확인하기 어려움
Expertise	Data Science, Machine Learning	Data Analysis

3. Why: Task Abstraction

	Task	Action+Target	Why	Outcome
Emily	인구 통계학적으로 top-rated 장르를 식별	Search Outliers	인구 통계학적 선호도를 이해	인구 통계별 콘텐츠 큐레이션의 정확도 향상
	시간에 따른 영화 장르, 감독, 배우의 평점 변화를 추적	Discover Trends	영화 인기의 변화하는 트렌드를 감지	추천 시스템의 동적 적응
John	영화 장르, 감독, 배우와 유저 그룹별 선호도 추적	Discover Trends	마케팅 캠페인을 최신 트렌드와 일치	타깃 광고의 효과 향상

4. How: Vis Idiom Design

1. Multi-line graph

y축은 년도별 평균 평점,

x축을 영화의 attribute으로 변경할 수 있음 (감독, 장르, 배우),

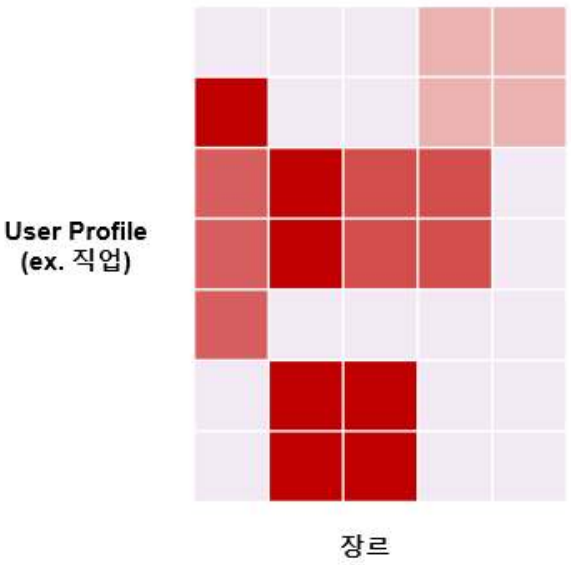
zoom하면 년도가 아닌 월 > 주 > 일 단위로 점차 scale이 세밀해짐

2. Heatmap

y축은 user의 profile (직업, 나이, 성별 등),

x축은 영화의 profile (장르, 감독, 배우 등)으로 변경할 수 있음.

tooltip하면 해당 칸의 유저 정보와 영화 정보가 display됨.

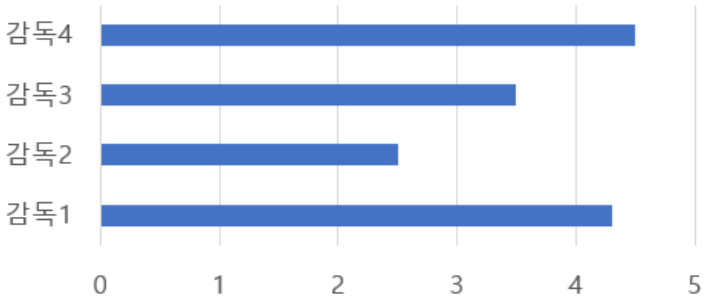


3. Bar graph

y축은 영화의 profile이며 (장르, 감독, 배우 등)로 변경할 수 있음,

x축은 전체 평균.

y축의 영화 profile을 변경하며 볼 수 있음.



5. Sketch

6. Implementation using Vega or D3.js

- 이름: Interactive Movie Trend Viualization tool
- 링크: https://bw-99.github.io/InfoViz2024/src/
- 스크린샷:

7. Case Study

- 1. John (Marketing Manager): 임팩트 있는 마케팅을 위한 차년도 인기 급상할 영화 예측
  - 올해 어떤 영화 장르가 가장 인기 있었는지를 3) Bar Chart를 통해 확인 -> Film-Noir
  - 1) Multi line graph를 통해 Film-Noir의 인기 추세를 확인 -> 올해 급격히 증가한 것을 확인
  - 내년 마케팅 캠패인 영화 대상을 Film-Noir 위주로 고려

- 2. John (Marketing Manager): 임팩트 있는 마케팅을 위한 차년도 인기 급감할 영화 예측
  - 전체 기간 (2000 ~ 2003) 동안 가장 인기 있었던 영화 장르를 2) Heatmap을 통해 확인 -> Short가 전반적으로 인기 많았음
  - 1) Multi line graph에서 short 검색하여 트랜드 확인 -> 급감한 것을 확인
  - 내년 마케팅 캠패인 영화 대상에서 Short는 배제

- 3. John (Marketing Manager): 전속 계약을 위한 배우 검색
  - 2) Heatmap을 통해 18세 이하의 사람들이 어떤 배우를 선호하는지 확인 -> Keiju Kobayashi

- 4. John (Marketing Manager): 전속 계약을 위한 감독 검색
  - 2) Heatmap을 통해 18세 이하의 사람들이 어떤 감독을 선호하는지 확인 -> Satyajit Ray
  - 1) Multi line graph에서 추세를 확인 -> 2001년을 기점으로 감소 중 -> 전속 계약 감독 후보에서 배제

- 5. Emily (Data Scientist): 시간에 따른 영화 장르 평균 평점의 변화를 추적
  - 관심 있는 영화 장르를 1) Multi line graph에 입력하여 추세 확인

- 6. Emily (Data Scientist): 인구 통계학적으로 top-rated된 장르를 식별
  - 2) Heatmap을 통해 연령, 직업, 성별에 따라 어떤 장르를 선호하는지 확인