데이터 분석, 무엇을, 어떻게

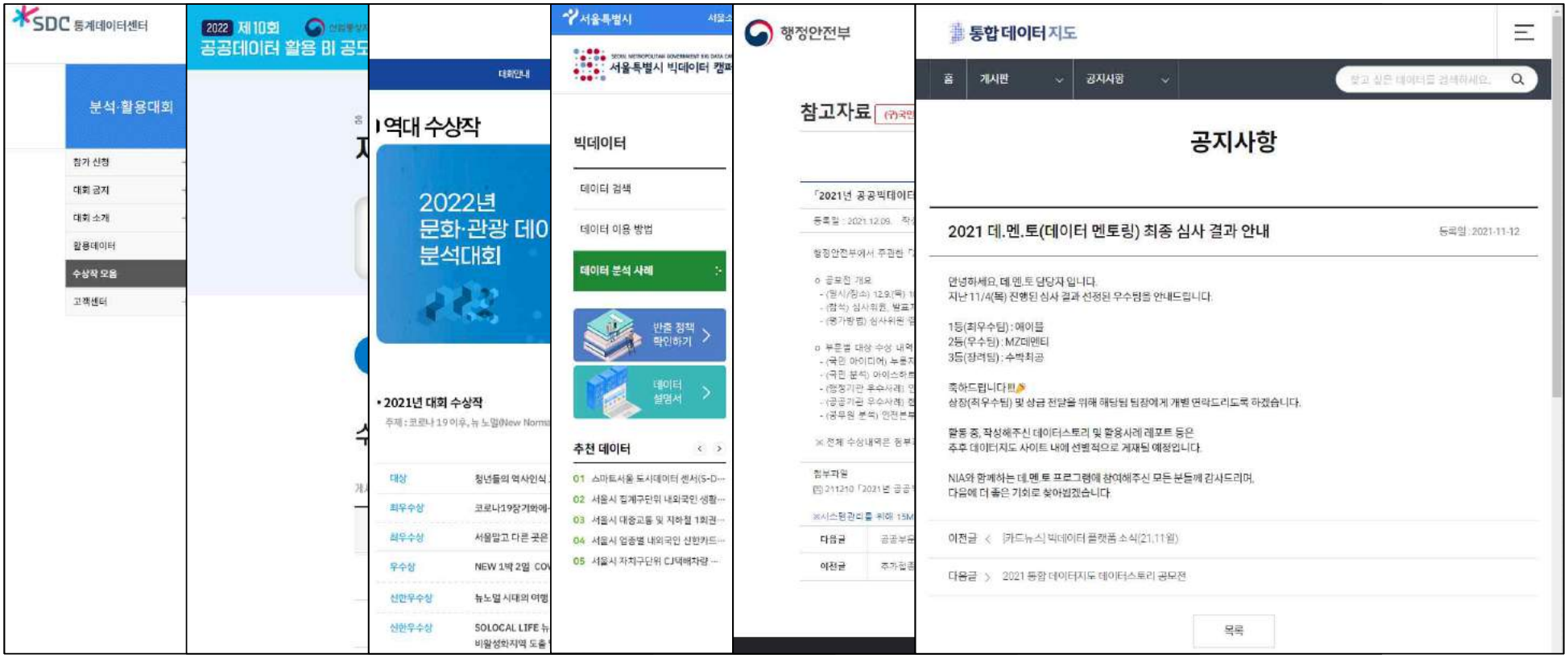
2022.07.21.

한국에너지기술연구원 이제현

데이터 분석 공모전 7건, 입상작 69건

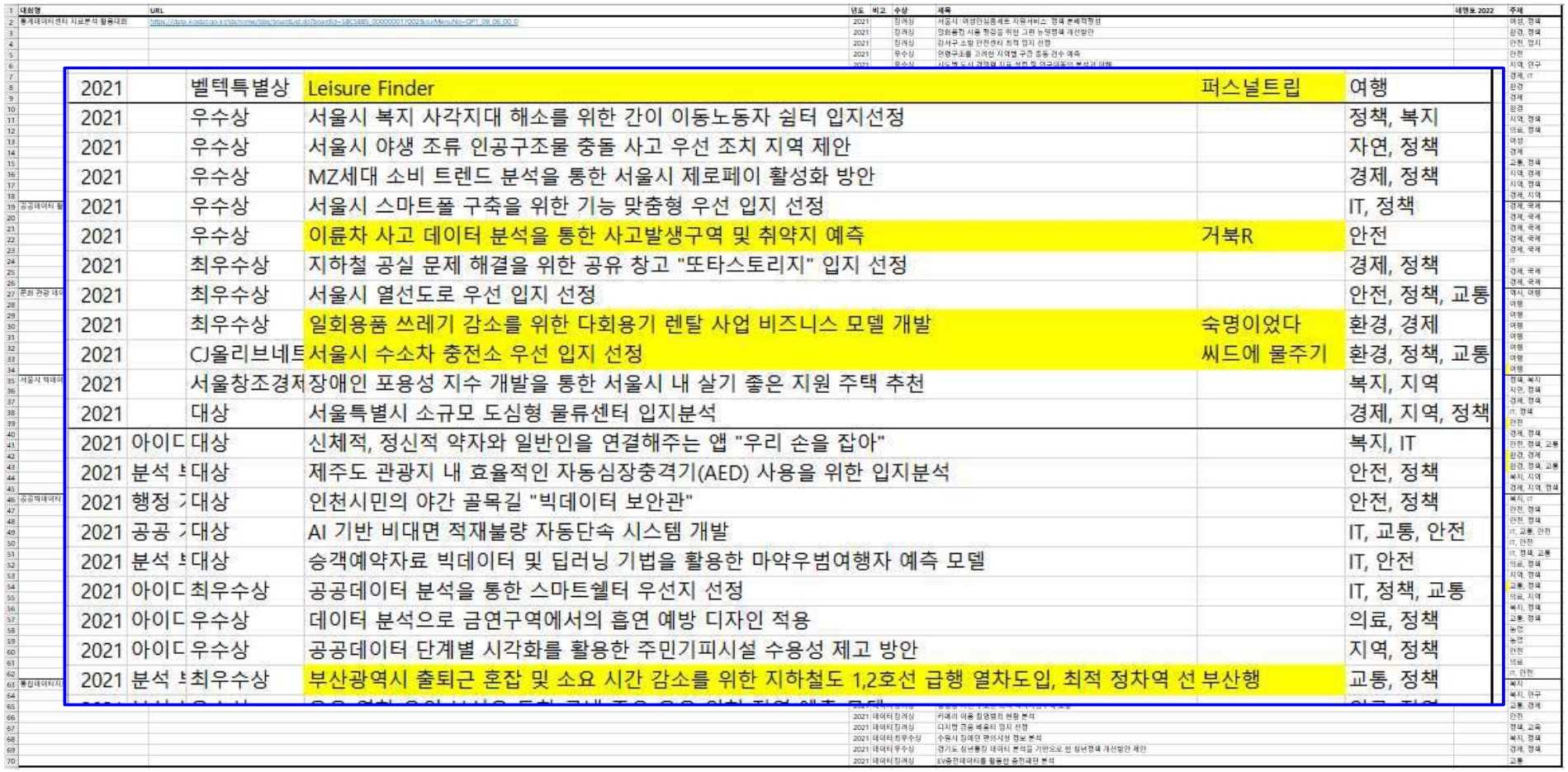
통계데이터센터 자료분석 활용대회　　공공데이터 활용 BI 공모전　　문화 관광 데이터 분석대회

서울시 빅데이터 캠퍼스 공모전　　공공빅데이터 분석 공모전　　통합데이터지도 데이터스토리 공모전, 데이터 멘토링



데이터 분석 공모전 7건, 입상작 69건

- 데이터 수집, 정리



2021		벨텍특별상	Leisure Finder	퍼스널트립	여행
2021		우수상	서울시 복지 사각지대 해소를 위한 간이 이동노동자 쉼터 입지선정		정책, 복지
2021		우수상	서울시 야생 조류 인공구조물 충돌 사고 우선 조치 지역 제안		자연, 정책
2021		우수상	MZ세대 소비 트렌드 분석을 통한 서울시 제로페이 활성화 방안		경제, 정책
2021		우수상	서울시 스마트폴 구축을 위한 기능 맞춤형 우선 입지 선정		IT, 정책
2021		우수상	이륜차 사고 데이터 분석을 통한 사고발생구역 및 취약지 예측	거북R	안전
2021		최우수상	지하철 공실 문제 해결을 위한 공유 창고 "또타스토리지" 입지 선정		경제, 정책
2021		최우수상	서울시 열선도로 우선 입지 선정		안전, 정책, 교통
2021		최우수상	일회용품 쓰레기 감소를 위한 다회용기 렌탈 사업 비즈니스 모델 개발	숙명이었다	환경, 경제
2021		CJ올리브네트	서울시 수소차 충전소 우선 입지 선정	씨드에 물주기	환경, 정책, 교통
2021		서울창조경저	장애인 포용성 지수 개발을 통한 서울시 내 살기 좋은 지원 주택 추천		복지, 지역
2021		대상	서울특별시 소규모 도심형 물류센터 입지분석		경제, 지역, 정책
2021	아이디	대상	신체적, 정신적 약자와 일반인을 연결해주는 앱 "우리 손을 잡아"		복지, IT
2021	분석 ㅏ	대상	제주도 관광지 내 효율적인 자동심장충격기(AED) 사용을 위한 입지분석		안전, 정책
2021	행정 ㅈ	대상	인천시민의 야간 골목길 "빅데이터 보안관"		안전, 정책
2021	공공 ㅈ	대상	AI 기반 비대면 적재불량 자동단속 시스템 개발		IT, 교통, 안전
2021	분석 ㅏ	대상	승객예약자료 빅데이터 및 딥러닝 기법을 활용한 마약우범여행자 예측 모델		IT, 안전
2021	아이디	최우수상	공공데이터 분석을 통한 스마트쉘터 우선지 선정		IT, 정책, 교통
2021	아이디	우수상	데이터 분석으로 금연구역에서의 흡연 예방 디자인 적용		의료, 정책
2021	아이디	우수상	공공데이터 단계별 시각화를 활용한 주민기피시설 수용성 제고 방안		지역, 정책
2021	분석 ㅏ	최우수상	부산광역시 출퇴근 혼잡 및 소요 시간 감소를 위한 지하철도 1,2호선 급행 열차도입, 최적 정차역 선	부산행	교통, 정책

데이터 분석 공모전 7건, 입상작 69건

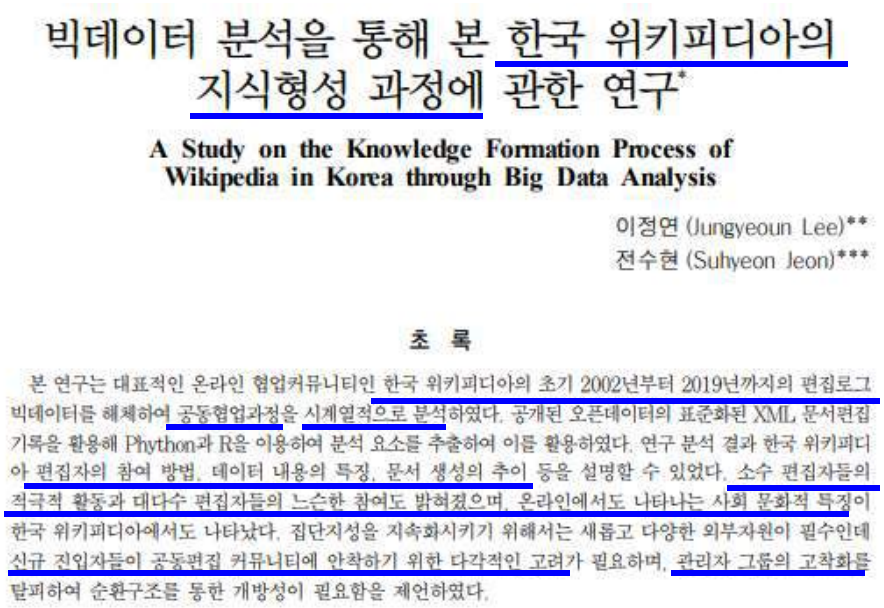
'서울시 "여성안심홈세트 지원서비스" 정책 분배적절성, 일회용컵 사용 절감을 위한 그린 뉴딜정책 개선방안, 성공적인 도시 재생 뉴딜사업을 위한 선정지표 개선 연구,
지역별 건강 격차에 따른 의료 자원 조정의 필요성 검토, 서울시 버스 혼잡도 예측 통한 다람쥐버스 신규 노선 제안, 신도시 타당성 요인 분석,
서울시 복지 사각지대 해소를 위한 간이 이동노동자 쉼터 입지선정, 서울시 야생 조류 인공구조물 충돌 사고 우선 조치 지역 제안, MZ세대 소비 트렌드 분석을 통한 서울시 제로페이 활성화 방안,
서울시 스마트폴 구축을 위한 기능 맞춤형 우선 입지 선정, 지하철 공실 문제 해결을 위한 공유 창고 "또타스토리지" 입지 선정, 서울시 열선도로 우선 입지 선정,
서울시 수소차 충전소 우선 입지 선정, 서울특별시 소규모 도심형 물류센터 입지분석, 제주도 관광지 내 효율적인 자동심장충격기(AED) 사용을 위한 입지분석,
인천시민의 야간 골목길 "빅데이터 보안관", 공공데이터 분석을 통한 스마트쉘터 우선지 선정, 데이터 분석으로 금연구역에서의 흡연 예방 디자인 적용,
공공데이터 단계별 시각화를 활용한 주민기피시설 수용성 제고 방안, 부산광역시 출퇴근 혼잡 및 소요 시간 감소를 위한 지하철도 1,2호선 급행 열차도입, 최적 정차역 선정,
시민 공원 이용 만족도 향상 및 효율적 민원 처리를 위한 공원 민원 빅데이터 분석, 신월-신정 다람쥐버스 도임, 디지털 금융 배움터 입지 선정, 수원시 장애인 편의시설 정보 분석,
경기도 청년통장 데이터 분석을 기반으로 한 청년정책 개선방안 제안'

'전통시장 DT 활용 방안, 인공지능을 활용한 가계금융건강검진, 빅데이터를 활용한 사업부지 맞춤형 컨설팅, 지역의 사회구조적 특성이 빈집 형성에 미친 영향 분석,
골목시장 방송 프로그램의 골목시장 활성화 효과 분석, 수출액 예측을 통한 수출 유망 국가와 품목 추천, 코트라 차년도 수출액 예측 과제,
클러스터링을 활용한 무역포트폴리오 다양화, KOTRA 수출 유망국가 추천, 경제적,
산업구조적, 문화적 요인을 기반으로 한 주요 국가의 한국 품목별 수입액 예측모형 개발: 한국의, 한국에 대한 문화적 요인을 중심으로, 분리학습 모델을 통한 국가별 품목별 수입액 예측 및 유망국가 추천,
한국 수입액 예측을 통한 유망 시장 탐색, MZ세대 소비 트렌드 분석을 통한 서울시 제로페이 활성화 방안, 지하철 공실 문제 해결을 위한 공유 창고 "또타스토리지" 입지 선정,
일회용품 쓰레기 감소를 위한 다회용기 렌탈 사업 비즈니스 모델 개발, 서울특별시 소규모 도심형 물류센터 입지분석,
통행량 기반 수도권 최적 택시사업구역 도출, 경기도 청년통장 데이터 분석을 기반으로 한 청년정책 개선방안 제안'

데이터 분석, 무엇을

데이터 분석, 왜

- 데이터 분석
  - Wikipedia: 유용한 정보를 발굴하고, 결론적인 내용을 알리며, 의사결정을 지원하는 것을 목표로 데이터를 정리, 변환, 모델링하는 과정



빅데이터 분석을 통해 본 한국 위키피디아의 지식형성 과정에 관한 연구*

A Study on the Knowledge Formation Process of Wikipedia in Korea through Big Data Analysis

이정연 (Jungyeoun Lee)**
전수현 (Suhyeon Jeon)***

초 록

본 연구는 대표적인 온라인 협업커뮤니티인 한국 위키피디아의 초기 2002년부터 2019년까지의 편집로그 빅데이터를 해체하여 공동협업과정을 시계열적으로 분석하였다. 공개된 오픈데이터의 표준화된 XML 문서편집 기록을 활용해 Phython과 R을 이용하여 분석 요소를 추출하여 이를 활용하였다. 연구 분석 결과 한국 위키피디아 편집자의 참여 방법, 데이터 내용의 특징, 문서 생성의 추이 등을 설명할 수 있었다. 소수 편집자들의 적극적 활동과 대다수 편집자들의 느슨한 참여도 밝혀졌으며, 온라인에서도 나타나는 사회 문화적 특징이 한국 위키피디아에서도 나타났다. 집단지성을 지속화시키기 위해서는 새롭고 다양한 외부자원이 필수인데 신규 진입자들이 공동편집 커뮤니티에 안착하기 위한 다각적인 고려가 필요하며, 관리자 그룹의 고착화를 탈피하여 순환구조를 통한 개방성이 필요함을 제언하였다.

분석 주제 "한국 위키피디아는 어떤 과정으로 형성되었나?"

분석 대상 "한국 위키피디아의 편집 로그 중 공동협업과정"

분석 방법 "시계열 분석"

분석 대상' "편집자의 참여 방법, 데이터 내용 특징, 문서 생성 추이"

유용한 정보 "적극적 소수와 느슨한 다수, 사회 문화적 특징"

의사결정 지원 "신규 진입자 지원, 관리자 그룹의 고착화 예방"

https://ko.wikipedia.org/
이정연, 전수현, "빅데이터 분석을 통해 본 한국 위키피디아의 지식형성 과정에 관한 연구", 정보관리학회지 37(2) 171-195, 2020

데이터 분석, 어떻게

데이터 수집 과정

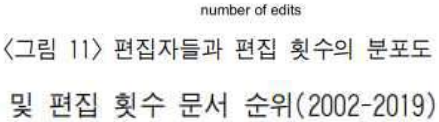
데이터 하나 골라 뜯어보기

한 발 물러서서 문서 전체를 바라보기

〈표 2〉 신규문서의 연도별 증가율

연도	신규 문서 수	전년대비 증가 수	증가율(%)	누적문서 수
2002	16			16
2003	454	438	▲2,738%	470
2004	4,823	4,369	▲962%	
2005	11,238	6,415	▲133%	
2006	13,989	2,751	▲24%	
2007	18,354	4,365	▲31%	
2008	33,952	15,598	▲85%	
2009	36,769	2,817	▲8%	
2010	31,745	-5,024	▼14%	
2011	33,990	2,245	▲7%	
2012	41,650	7,660	▲23%	
2013	31,448	-10,202	▼24%	
2014	37,187	5,739	▲18%	
2015	39,697	2,510	▲7%	
2016	32,417	-7,280	▼18%	
2017	37,903	5,486	▲17%	
2018	34,648	-3,255	▼9%	
2019	34,667	19	▲0%	
총 합계	474,947			

관점 변경: 편집자 바라보기



〈그림 11〉 편집자들과 편집 횟수의 분포도

〈표 5〉 한국위키피디아 편집자 최다 참여 및 편집 횟수 문서 순위(2002-2019)

순위	TITLE	편집자 수	편집시작연도	최근 편집연도	순위	TITLE	편집횟수	편집시작연도	최근편집연도
1	대한민국	1,870	2003	2019	1	개그콘서트	15,049	2008	2019
2	개그콘서트	1,674	2008	2019	2	박정희	7,335	2004	2019
3	조선민주주의인민공화국	1,479	2003	2019	3	미스터리음악쇼 복면가왕	6,950	2015	2019
4	박정희	1,419	2004	2019	4	대한민국	6,934	2003	2019
5	소녀시대	1,411	2007	2019	5	대한민국 축구 국가대표팀	6,592	2005	2019
6	1박2일	1,352	2008	2019	6	FC 서울	6,571	2005	2019
7	뮤직뱅크	1,236	2008	2019	7	1박 2일	6,326	2008	2019
8	노무현	1,225	2003	2019	8	소녀시대	6,133	2007	2019
9	대한민국축구국가대표팀	1,184	2005	2019	9	뮤직뱅크	6,017	2008	2019
10	자유한국당	1,128	2004	2019	10	도전! 골든벨의 에피소드 목록	5,354	2013	2019

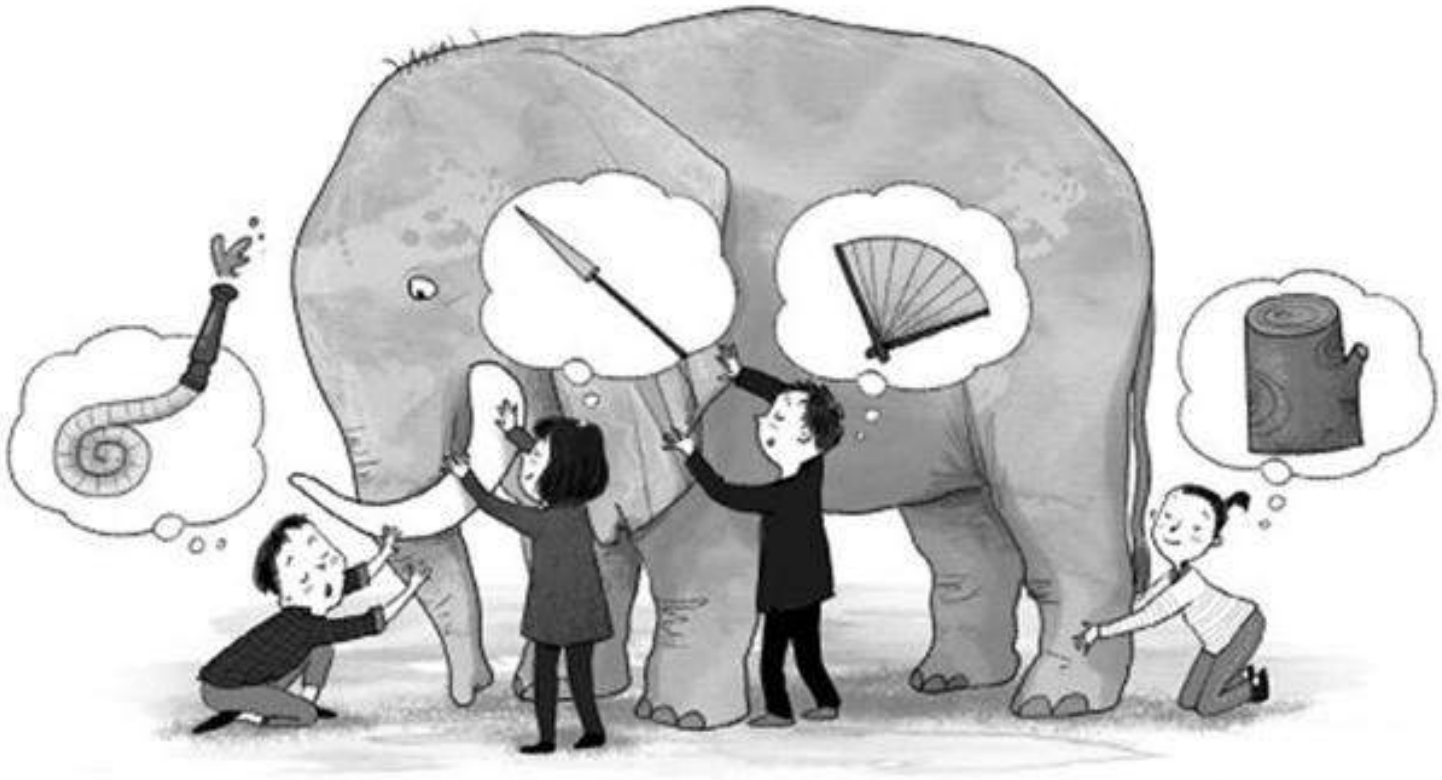
결론

① 위키피디아는 잘 유지, 관리되고 있다.
② 한국 위키피디아는 신규 문서가 꾸준히 증가하고 있으며 ,
특히 연예, 오락, 스포츠, 정치, 사회 업데이트가 활발하다
③ 비로그인 그룹의 편집이 로그인 그룹의 4배에 달한다.
이런 추세는 2010년경부터 두드러진다.
④ 극소수의 적극적 편집자 + 절대 다수의 1~2회 편집자가 있다.
⑤ 관리자는 22인이며 한국식 위계질서(가입 우선순위)가 관찰된다
⑥ 신규 진입이 꾸준하나 활성 편집자로 안착되지 않고 있다.

이정연, 전수현, “빅데이터 분석을 통해 본 한국 위키피디아의 지식형성 과정에 관한 연구”, 정보관리학회지 37(2) 171-195, 2020

Exploratory Data Analysis

- 목표 : 데이터 파악



https://www.mk.co.kr/today-paper/view/2020/4663062/

데이터 분석, 왜

- Q. “관리자나 고인물한테 물어보면 다 아는 것 아니예요?”
- A1.
- A2.
- A3.

https://e.kakao.com/t/the-gap-between-talk-and-action-3

데이터 분석, 무엇을

- 데이터 분석
  - Wikipedia: 유용한 정보를 발굴하고, 결론적인 내용을 알리며, 의사결정을 지원하는 것을 목표로 데이터를 정리, 변환, 모델링하는 과정

→ 유용한 정보가 나올 만한 것

→ 결론적인 내용이 나올 만한 것

→ 의사 결정에 지원이 될 만한 것

정량적으로 측정할 수 있는 것 = 객관적으로 전달할 수 있는 것

객관적으로 전달할 수 없는 것 = 감각에 의존하는 것 = 관능 평가

"if you can't measure it, you can't manage it"

- Peter Drucker

"It is wrong to suppose that if you can't measure it, you can't manage it – a costly myth."

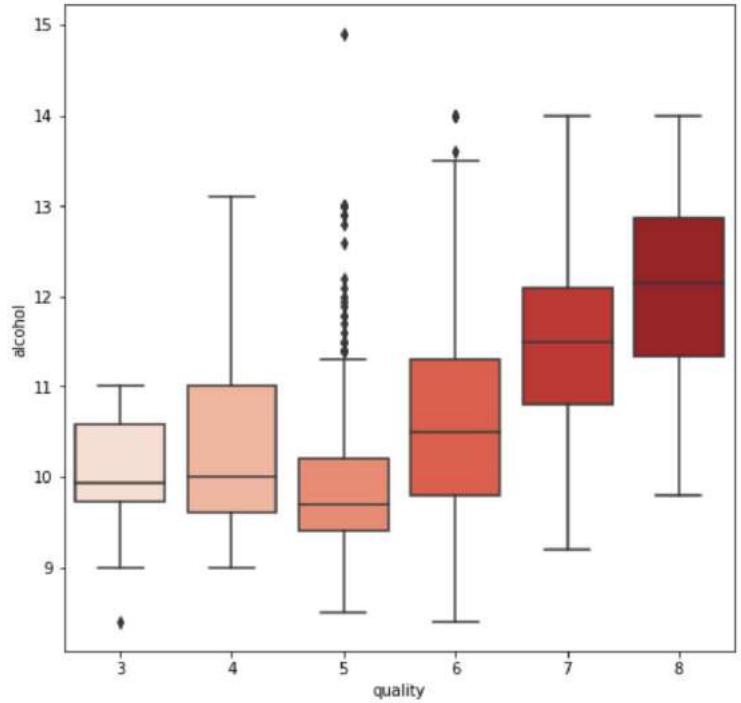
- W. Edwards Deming

데이터 분석, 무엇을

- 객관적으로 전달할 수 없는 것 = 감각에 의존하는 것 = 관능 평가

"It is wrong to suppose that if you can't measure it, you can't manage it – a costly myth."

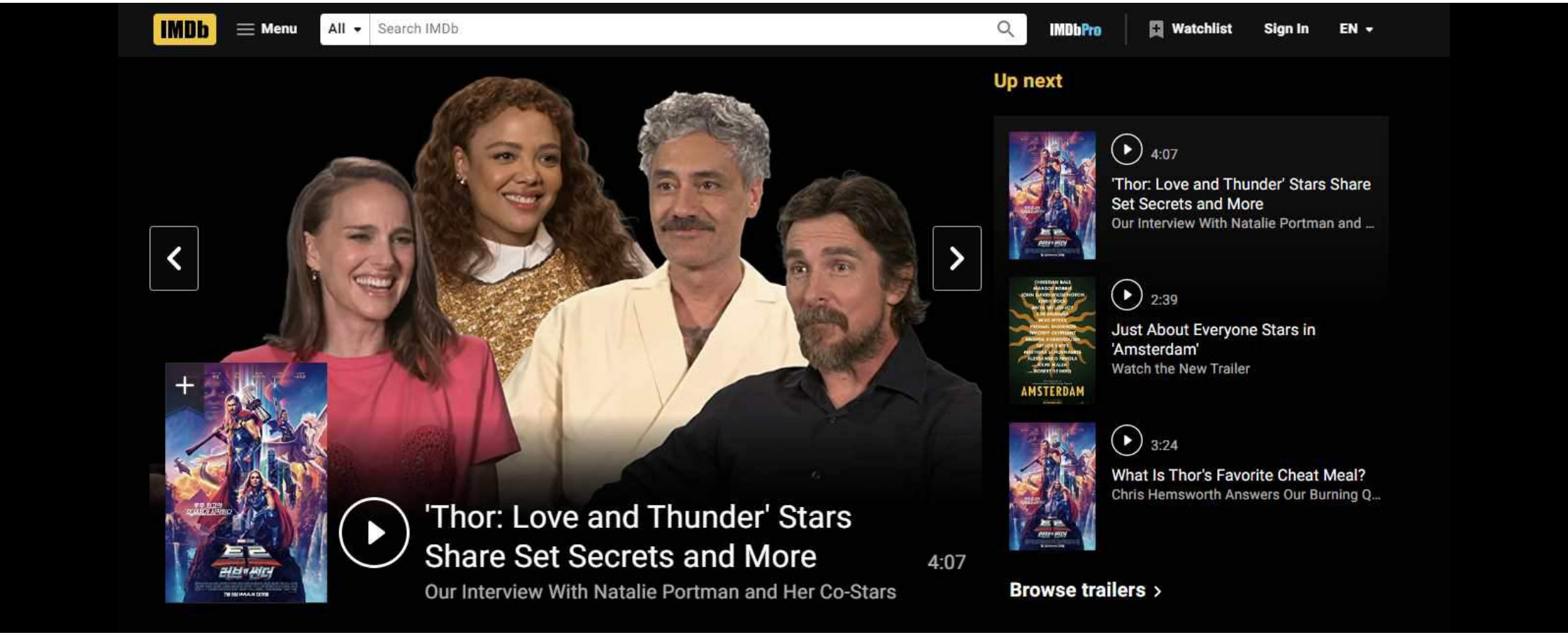
- W. Edwards Deming



https://psuwineandgrapes.wordpress.com/2015/11/27/whats-wrong-with-my-wine-bringing-you-ways-to-improve-the-quality-of-your-wine-by-minimizing-the-effects-of-wine-defects/
https://towardsdatascience.com/what-makes-a-wine-good-ea370601a8e4

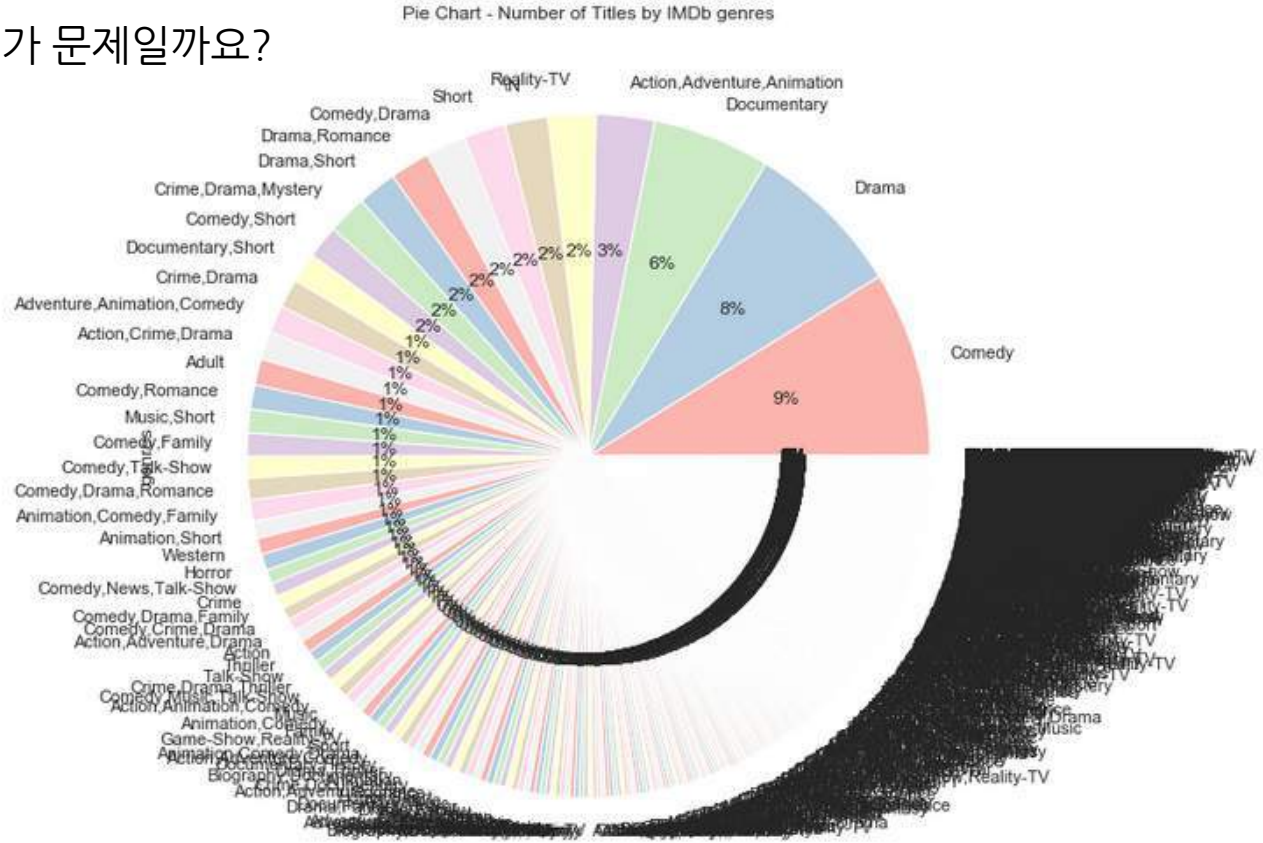
예제 : 영화 데이터

- IMDB dataset



데이터 분석 예시: 장르

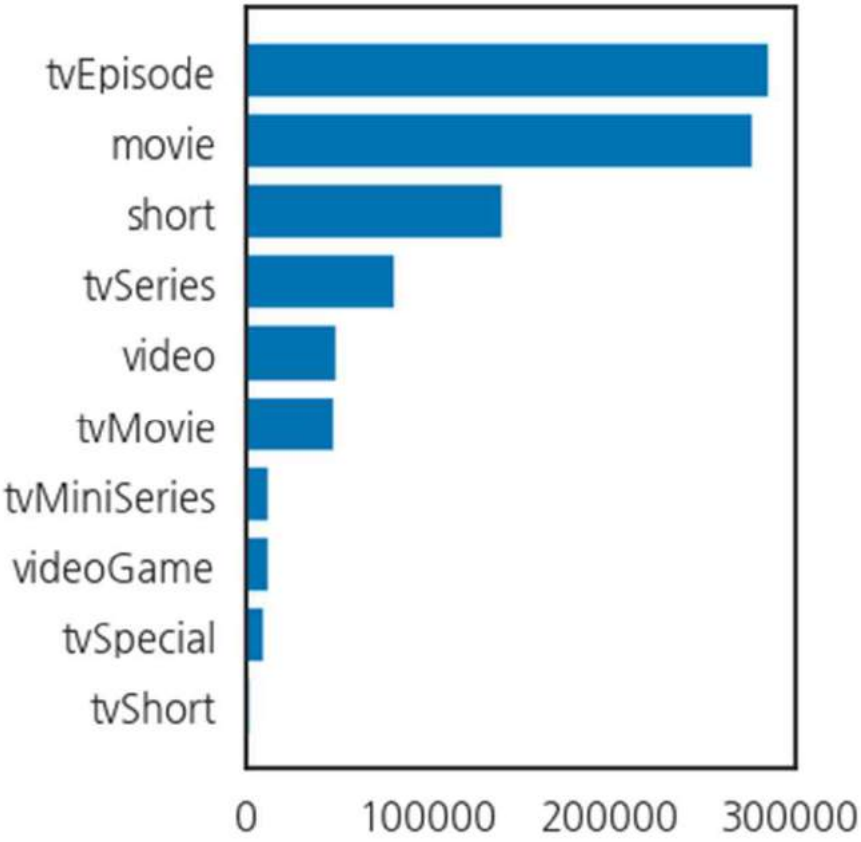
- 이거 제대로 된 걸까요?
- 잘못됐다면 어디가 문제일까요?

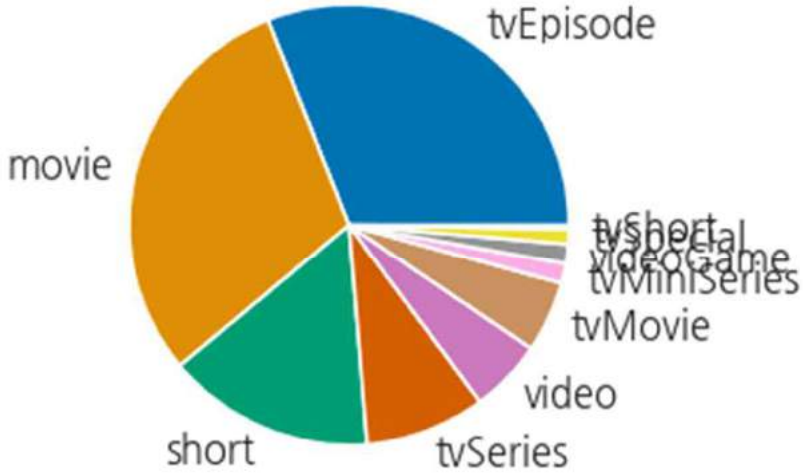


https://bit.ly/3cdsVj2

데이터 분석 예시: 장르 < 유형

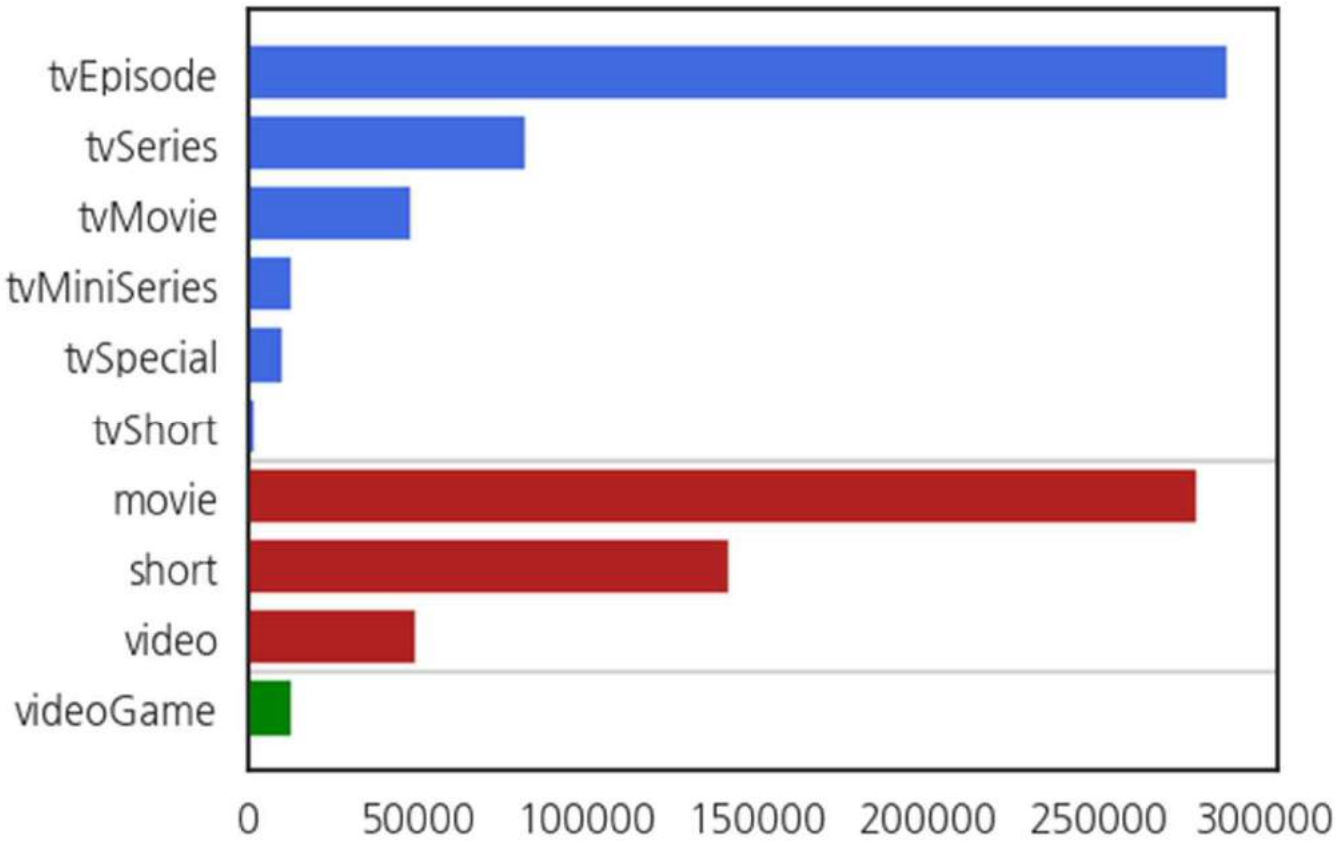
- 작은 분류로 들어가기 전에 큰 분류부터 : titleType





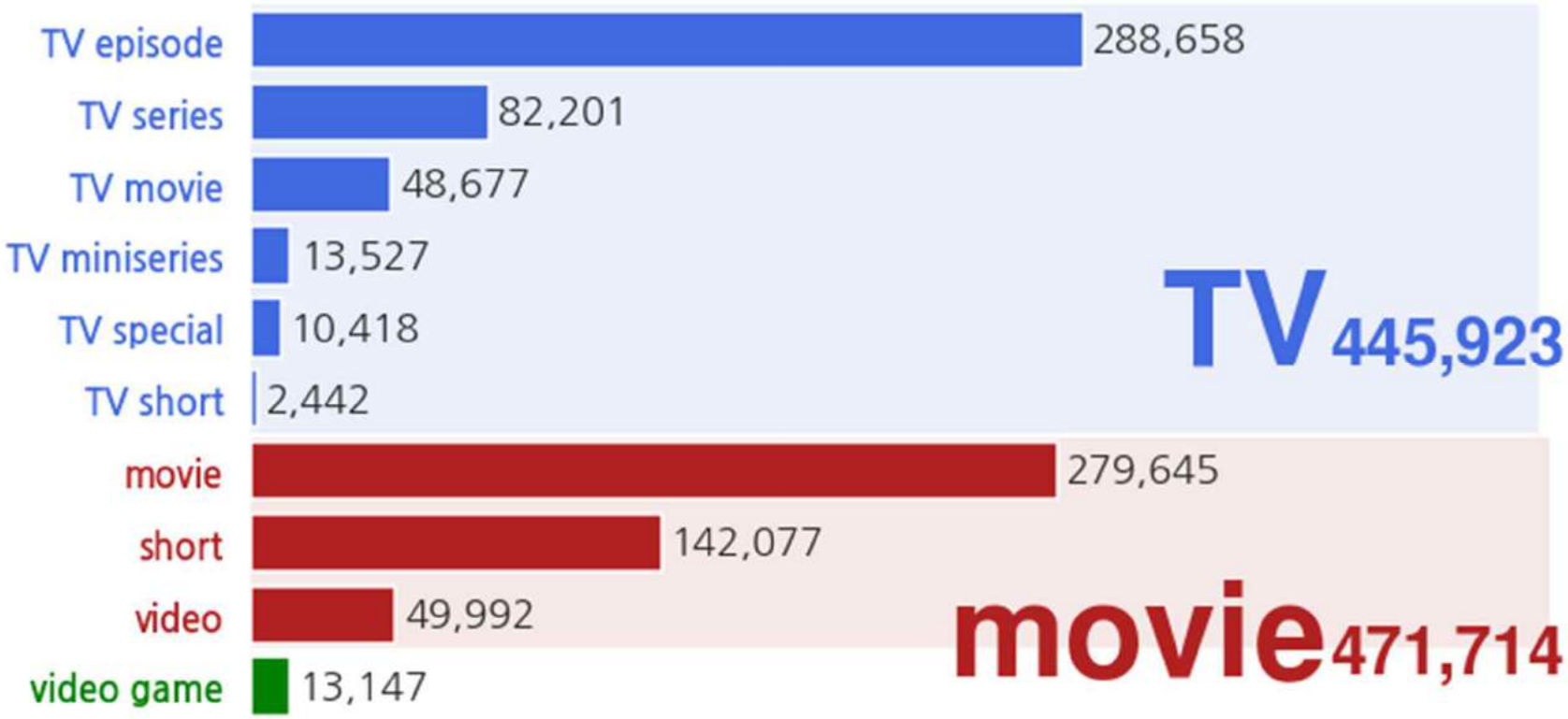
데이터 분석 예시: 장르 < 유형

- 작은 분류로 들어가기 전에 큰 분류부터 : titleType



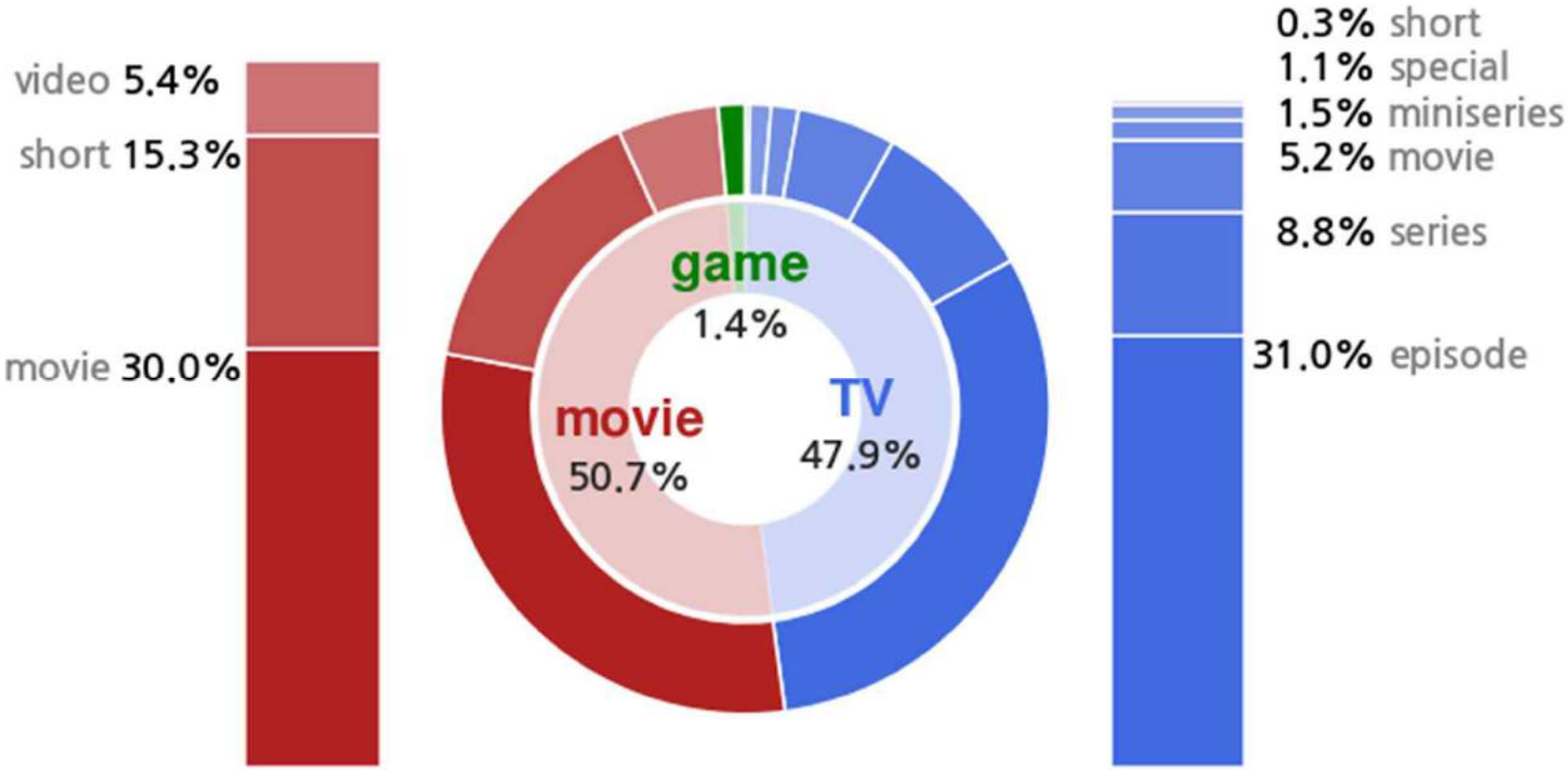
데이터 분석 예시: 장르 < 유형

- 기왕이면 깔끔(?)하게



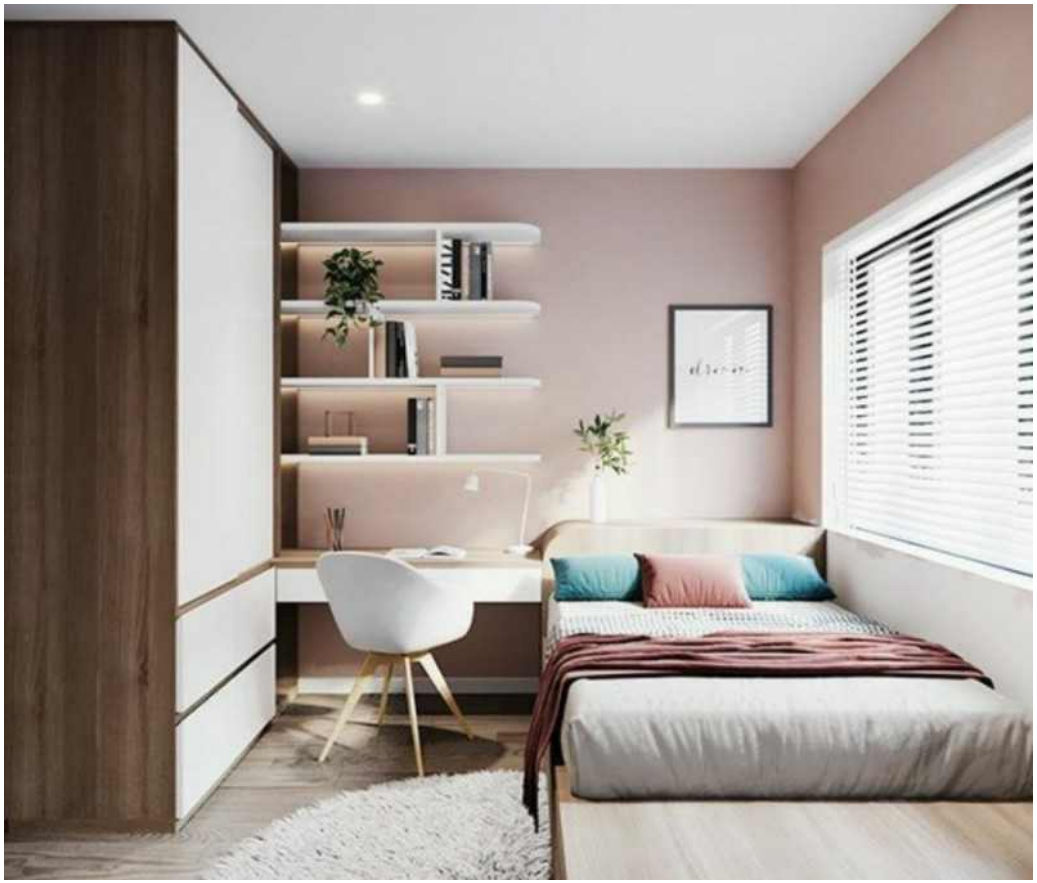
데이터 분석 예시: 장르 < 유형

- 기왕이면 깔끔(?)하게



시각화 투자 에너지

- 방 정리 상태



https://m.blog.naver.com/livinghyun22/221956042101

https://m.cafe.daum.net/topstoncomputer/JlDJ/12

데이터 직접 확인: sampling

- 데이터 내용을 직접 눈으로 확인
  - 2000년 이후, 장르별 투표자수 top 10 → “내가 아는 작품 찾기”

```
df_region_nn = df_region.loc[df_region["startYear"] != "\\N"]
for titleType in df_region["titleType"].unique():
    titles = df_region_nn.loc[df_region_nn["startYear"].astype(int) >= 2000].loc[df_region["titleType"]==titleType].sort_values("numVotes", ascending=False)["originalTitle"].iloc[:5].values
    years = df_region_nn.loc[df_region_nn["startYear"].astype(int) >= 2000].loc[df_region["titleType"]==titleType].sort_values("numVotes", ascending=False)["startYear"].iloc[:5].values
    print(f"# {titleType}: {dict(zip(titles, years))}\n")
```

```
# short: {'Kung Fury': '2015', 'Hotel Chevalier': '2007', 'Paperman': '2012', 'For the Birds': '2000', 'Piper': '2016'}

# movie: {'The Dark Knight': '2008', 'Inception': '2010', 'The Lord of the Rings: The Fellowship of the Ring': '2001', 'The Lord of the Rings: The Return of the King': '2003', 'Interstellar': '2014'}

# tvEpisode: {'The Iron Throne': '2019', 'The Long Night': '2019', 'Battle of the Bastards': '2016', 'The Bells': '2019', 'Ozymandias': '2013'}

# tvSeries: {'Game of Thrones': '2011', 'Breaking Bad': '2008', 'Stranger Things': '2016', 'The Walking Dead': '2010', 'Sherlock': '2010'}

# tvShort: {'Toy Story of Terror': '2013', 'Shrek the Halls': '2007', 'Toy Story That Time Forgot': '2014', 'Ice Age: A Mammoth Christmas': '2011', 'Robot Chicken: Star Wars': '2007'}

# tvMovie: {'High School Musical': '2006', 'High School Musical 2': '2007', 'Sharknado': '2013', 'The Normal Heart': '2014', 'Camp Rock': '2008'}

# tvMiniSeries: {'Chernobyl': '2019', 'Band of Brothers': '2001', "The Queen's Gambit": '2020', 'WandaVision': '2021', 'The Haunting of Hill House': '2018'}

# tvSpecial: {'Friends: The Reunion': '2021', 'Bo Burnham: Inside': '2021', 'Harry Potter 20th Anniversary: Return to Hogwarts': '2022', 'Death to 2020': '2020', 'Dave Chappelle: Sticks & Stones': '20

# video: {'The Animatrix': '2003', 'Band Camp': '2005', 'The Naked Mile': '2006', 'Beta House': '2007', 'Batman: Under the Red Hood': '2010'}

# videoGame: {'The Last of Us': '2013', 'Grand Theft Auto V': '2013', 'Grand Theft Auto: San Andreas': '2004', 'Grand Theft Auto IV': '2008', 'Red Dead Redemption II': '2018'}
```

데이터 직접 확인 : sampling

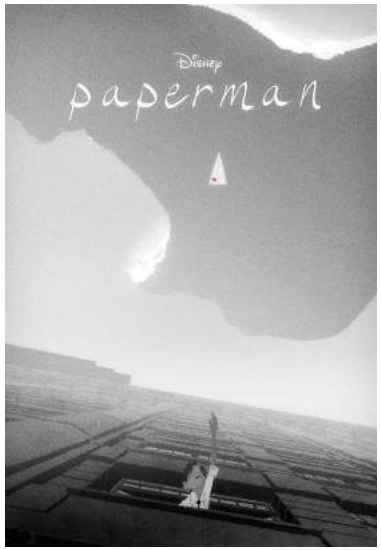
- “내가 알 만한 작품 찾기”
  - 2000년 이후, 장르별 투표자수 top 5

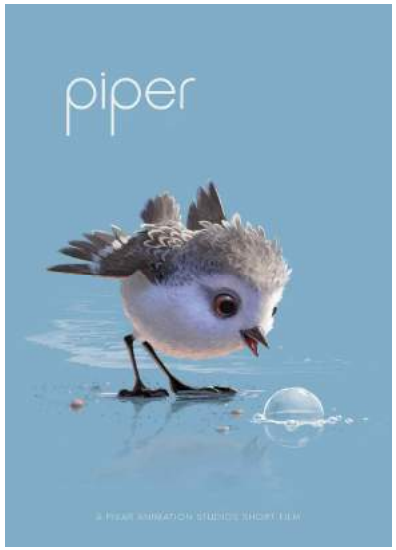
videoGame: {'The Last of Us': '2013', 'Grand Theft Auto V': '2013', 'Grand Theft Auto: San Andreas': '2004', 'Grand Theft Auto IV': '2008', 'Red Dead Redemption II': '2018'}

데이터 직접 확인 : sampling

- “내가 알 만한 작품 찾기”
  - 2000년 이후, 장르별 투표자수 top 5

short: {'Kung Fury': '2015', 'Hotel Chevalier': '2007', 'Paperman': '2012', 'For the Birds': '2000', 'Piper': '2016'}

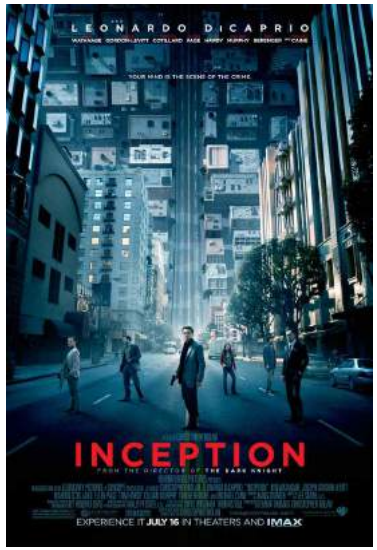




데이터 직접 확인 : sampling

- “내가 알 만한 작품 찾기”
  - 2000년 이후, 장르별 투표자수 top 5

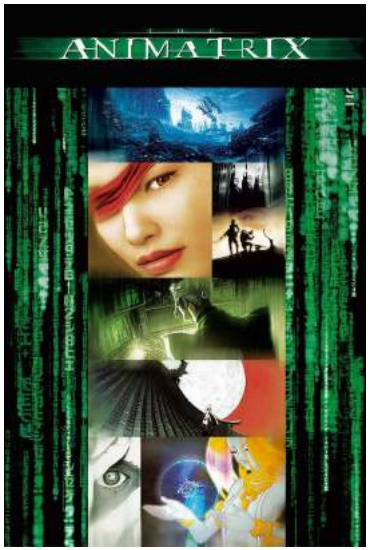
movie: {'The Dark Knight': '2008', 'Inception': '2010', 'The Lord of the Rings: The Fellowship of the Ring': '2001', 'The Lord of the Rings: The Return of the King': '2003', 'Interstellar': '2014'}

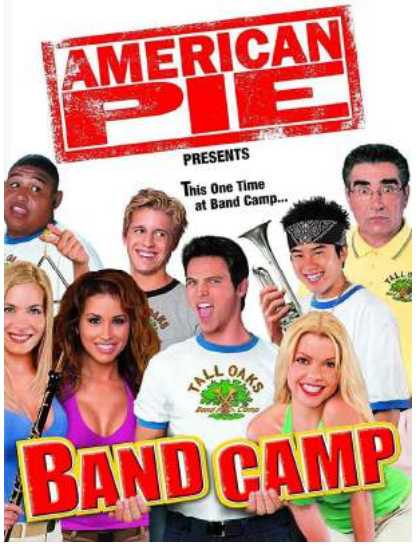


데이터 직접 확인 : sampling

- “내가 알 만한 작품 찾기”
  - 2000년 이후, 장르별 투표자수 top 5

video: {'The Animatrix': '2003', 'Band Camp': '2005', 'The Naked Mile': '2006', 'Beta House': '2007', 'Batman: Under the Red Hood': '2010'}





데이터 직접 확인 : sampling

- **“내가 알 만한 작품 찾기”**
  - 2000년 이후, 장르별 투표자수 top 5

tvSeries: {'Game of Thrones': '2011', 'Breaking Bad': '2008', 'Stranger Things': '2016', 'The Walking Dead': '2010', 'Sherlock': '2010'}

데이터 직접 확인 : sampling

- **“내가 알 만한 작품 찾기”**
  - 2000년 이후, 장르별 투표자수 top 5

tvEpisode: {'The Iron Throne': '2019', 'The Long Night': '2019', 'Battle of the Bastards': '2016', 'The Bells': '2019', 'Ozymandias': '2013'}





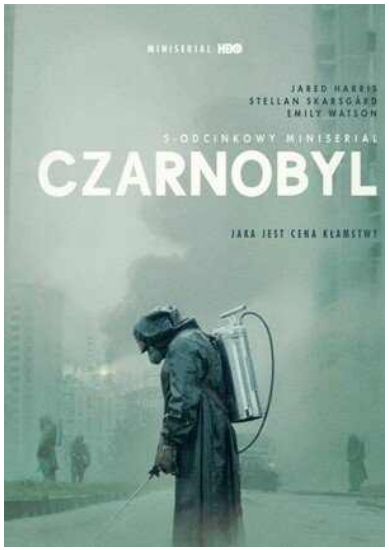
Game of Thrones

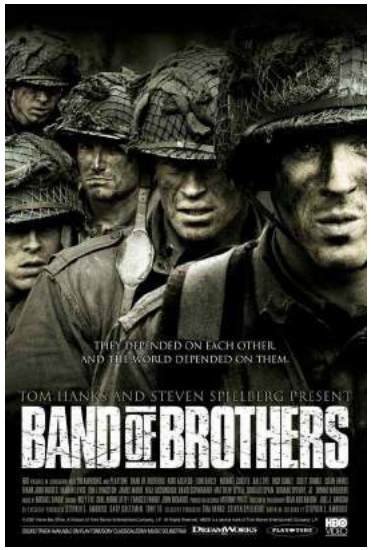
Breaking Bad

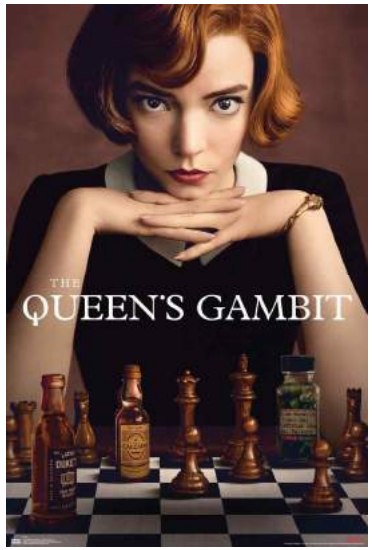
데이터 직접 확인 : sampling

- “내가 알 만한 작품 찾기”
  - 2000년 이후, 장르별 투표자수 top 5

tvMiniSeries: {'Chernobyl': '2019', 'Band of Brothers': '2001', "The Queen's Gambit": '2020', 'WandaVision': '2021', 'The Haunting of Hill House': '2018'}



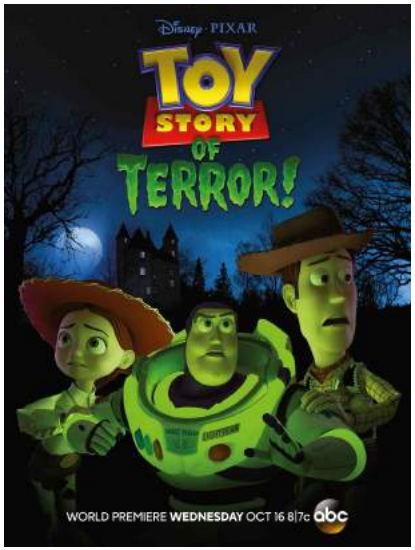


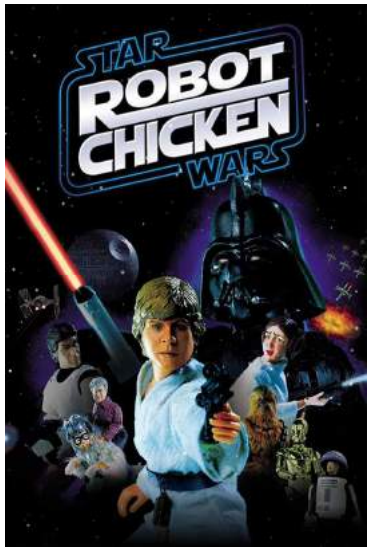


데이터 직접 확인 : sampling

- “내가 알 만한 작품 찾기”
  - 2000년 이후, 장르별 투표자수 top 5

tvShort: {'Toy Story of Terror': '2013', 'Shrek the Halls': '2007', 'Toy Story That Time Forgot': '2014', 'Ice Age: A Mammoth Christmas': '2011', 'Robot Chicken: Star Wars': '2007'}

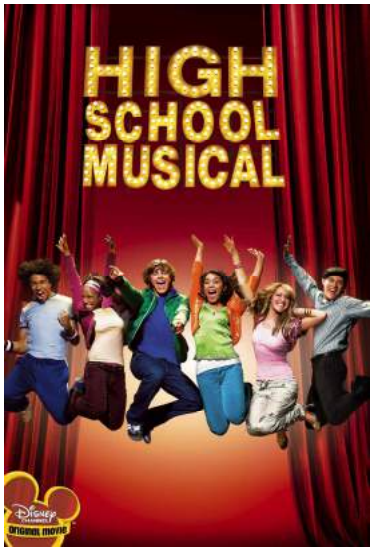


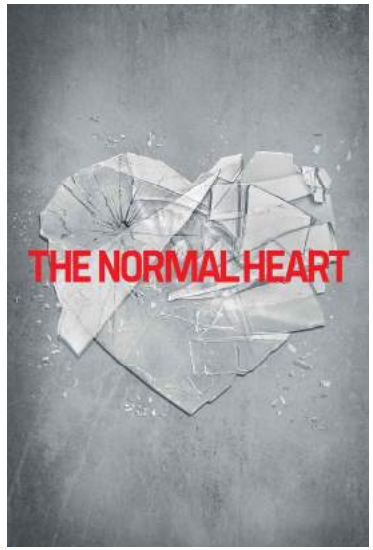


데이터 직접 확인 : sampling

- “내가 알 만한 작품 찾기”
  - 2000년 이후, 장르별 투표자수 top 5

tvMovie: {'High School Musical': '2006', 'High School Musical 2': '2007', 'Sharknado': '2013', 'The Normal Heart': '2014', 'Camp Rock': '2008'}

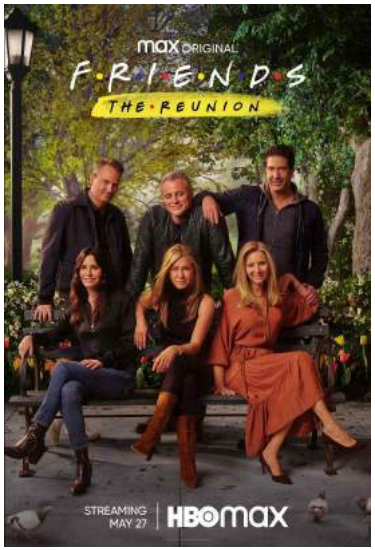


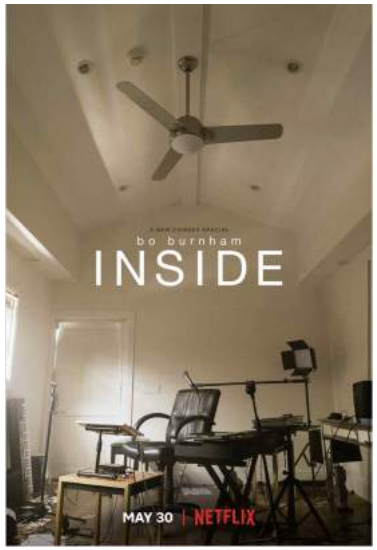


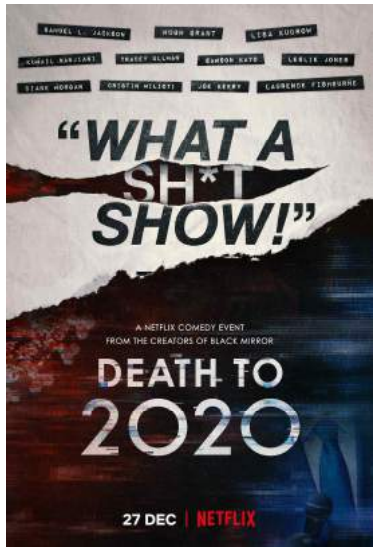
데이터 직접 확인 : sampling

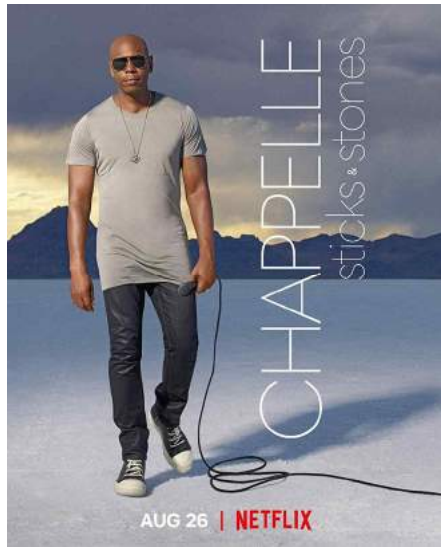
- “내가 알 만한 작품 찾기”
  - 2000년 이후, 장르별 투표자수 top 5

tvSpecial: {'Friends: The Reunion': '2021', 'Bo Burnham: Inside': '2021', 'Harry Potter 20th Anniversary: Return to Hogwarts': '2022', 'Death to 2020': '2020', 'Dave Chappelle: Sticks & Stones': '2019'}





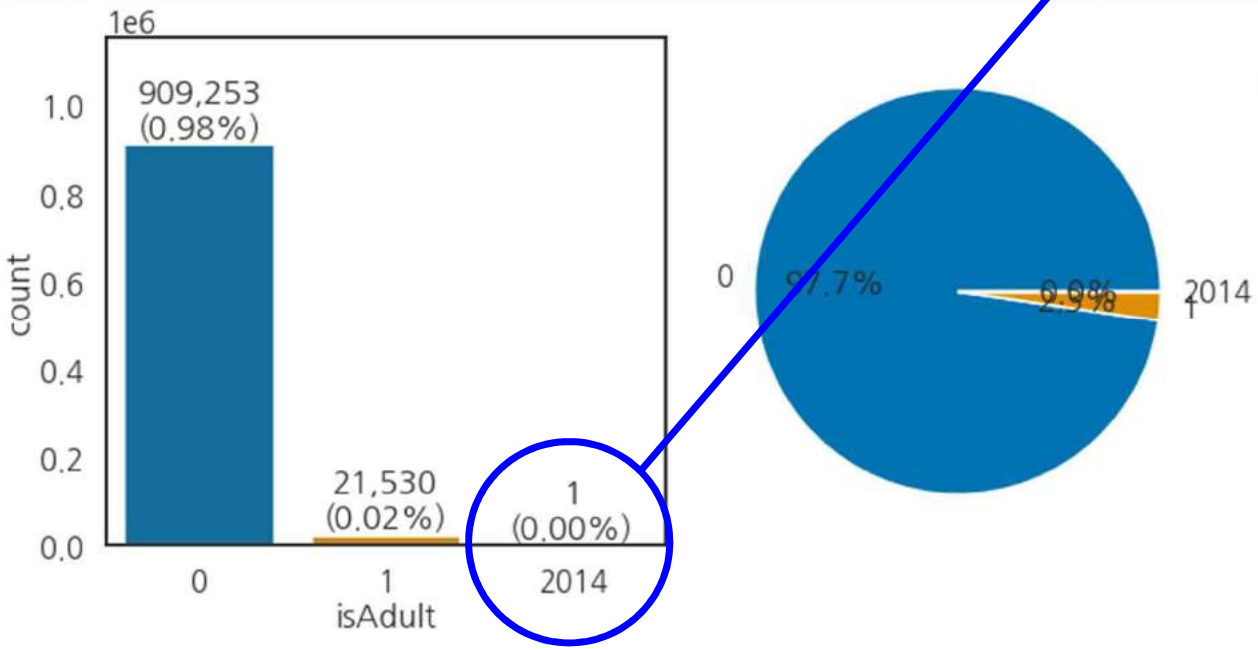


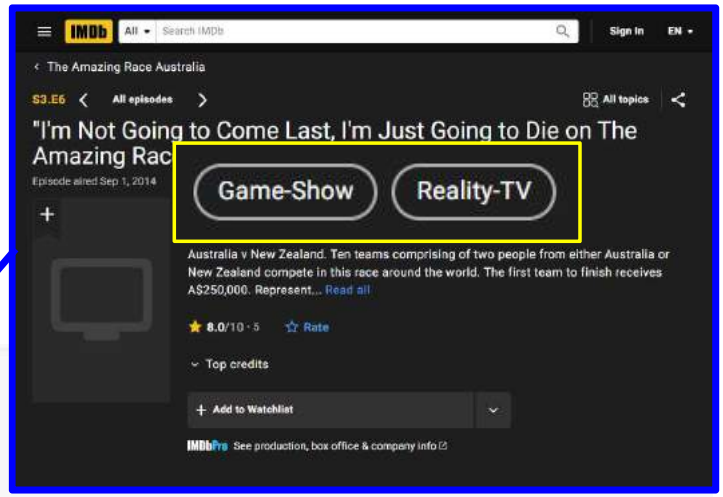


데이터 직접 확인 : isAdult

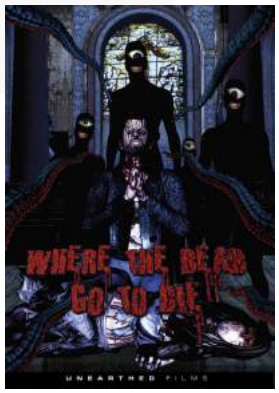
- 성인용 수, 비율

```
1 fig, axs = plt.subplots(ncols=2, figsize=(10, 5), constrained_layout=True)
2
3 # count
4 sns.countplot(x="isAdult", data=df_region, ax=axs[0])
5 offset = 2e4
```





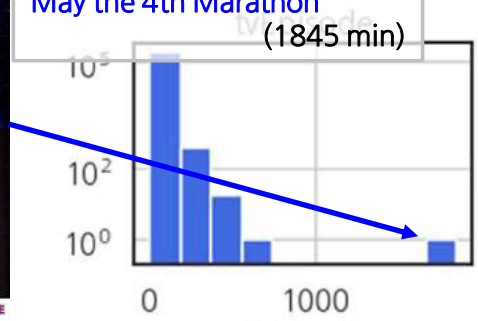
isAdult: {'Pirates': '2005', "Pirates II: Stagnetti's Revenge": '2008', '1 Night in Paris': '2004', 'Inran naru ichizoku: Dai-ni-shô - Zetsurin no hate ni': '2004', 'Where the Dead Go to Die': '2012'}

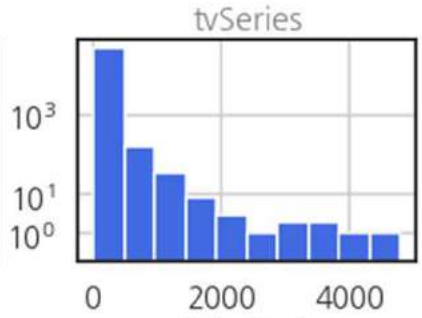


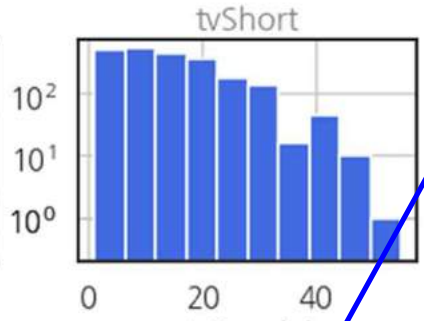
데이터 직접 확인 : runtimeMinutes

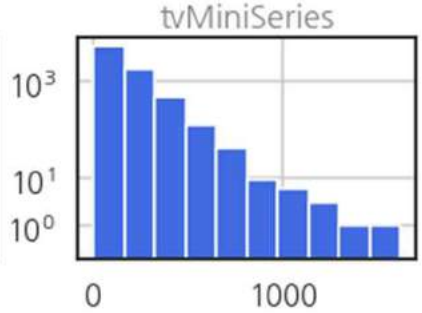
- **변수 이름은 가급적 바꾸지 말 것 :** 지나친 약어 등으로 알아보기 힘들지 않으면 데이터를 그대로 인지하기
- TV : Series, MiniSeries는 길어야 정상, Shorts는 짧아야 정상.
  - 그렇다면 유별나게 긴 것들은?

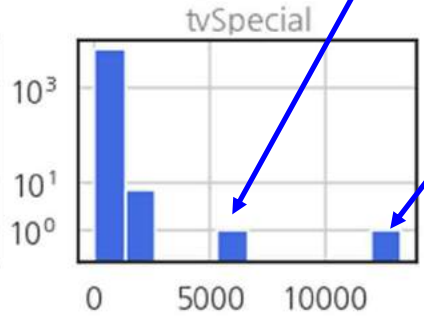
The George Lucas Talk Show
May the 4th Marathon
(1845 min)











24H Europe:
The Next
Generation
(1440 min)

24H Berlin
(1440 min)

Johann Wolfgang von
Goethe: Faust II
(1320 min)

데이터 직접 확인 : runtimeMinutes

- **변수 이름은 가급적 바꾸지 말 것 :** 지나친 약어 등으로 알아보기 힘들지 않으면 데이터를 그대로 인지하기
- TV : Series, MiniSeries는 길어야 정상, Shorts는 짧아야 정상.
  - 그렇다면 유별나게 긴 것들은? 확인하자! **일일이.**

```
[54] df_tmp.loc[df_tmp["titleType"]=="tvEpisode"].sort_values("runtimeMinutes", ascending=False)[:1]
```

	tconst	titleType	primaryTitle	originalTitle	isAdult	startYear	endYear	runtimeMinutes	genres	averageRating	numVotes	title	region	source
465544	tt12298094	tvEpisode	The George Lucas Talk Show May the 4th Marathon	May the 4th Marathon	0	2020	₩N	1845.0	Comedy,Talk-Show	9.6	54	[May the 4th Marathon, The George Lucas Talk S...	[₩N, US, ₩N]	tv

```
[55] df_tmp.loc[df_tmp["titleType"]=="tvMovie"].sort_values("runtimeMinutes", ascending=False)[:3]
```

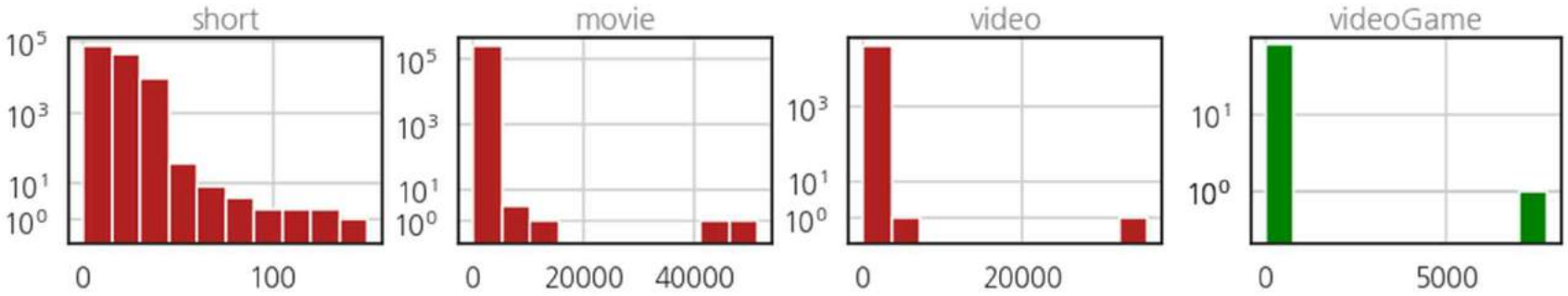
	tconst	titleType	primaryTitle	originalTitle	isAdult	startYear	endYear	runtimeMinutes	genres	averageRating	numVotes	title	region	source
900167	tt8539844	tvMovie	24H Europe: The Next Generation	24H Europe: The Next Generation	0	2019	₩N	1440.0	Documentary	6.3	20	[24h Europe: We Are the Future, 24H Europe: Th...	[₩N, DE, ₩N, DE]	tv
483835	tt1291623	tvMovie	24 Hours Berlin	24h Berlin - Ein Tag im Leben	0	2009	₩N	1440.0	Documentary	7.9	115	[24h Berlin - Ein Tag im Leben, 24 Hours Berli...	[DE, XWW, GR, ₩N]	tv
170437	tt0282642	tvMovie	Johann Wolfgang von Goethe: Faust II	Johann Wolfgang von Goethe: Faust II	0	2001	₩N	1320.0	Drama	8.2	46	[Johann Wolfgang von Goethe: Faust II, Johann ...	[₩N, DE]	tv

```
[56] df_tmp.loc[df_tmp["titleType"]=="tvSpecial"].sort_values("runtimeMinutes", ascending=False)[:2]
```

	tconst	titleType	primaryTitle	originalTitle	isAdult	startYear	endYear	runtimeMinutes	genres	averageRating	numVotes	title	region	source
446884	tt11707418	tvSpecial	Svalbard Minute by Minute	Svalbard minutt for minutt	0	2020	₩N	13319.0	Adventure,Documentary	8.1	21	[Svalbard Minute by Minute, Svalbard minutt fo...	[US, NO, XWW, GB, ₩N]	tv
867801	tt7357138	tvSpecial	Katy Perry Live: Witness World Wide	Katy Perry Live: Witness World Wide	0	2017	₩N	5760.0	Music,Reality-TV	2.8	181	[Katy Perry Live: Witness World Wide, Katy Per...	[₩N, US]	tv

데이터 직접 확인 : runtimeMinutes

- **movie, game** : movie, video 길이가 너무 길면 이상, Shorts는 짧아야 정상.



Big Chungus (2018)
(7770 min = 5.4 days)

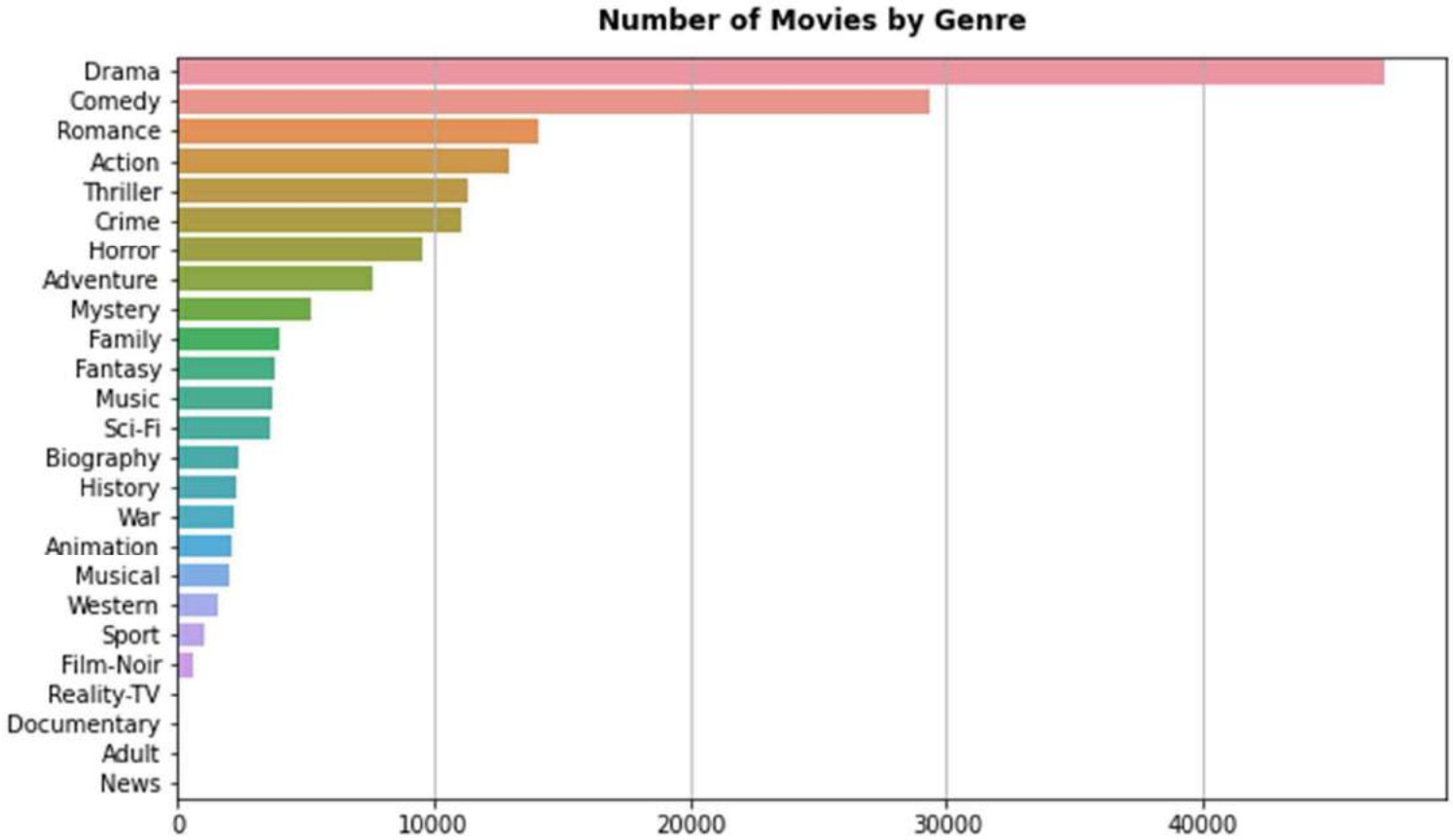
Ambiancé (2020)
(43200 min = 720 hour = 30 days)

Logistics (2012)
(51420 min = 857 hour = 35.7 days)

The Longest Video on YouTube: 596.5 Hours (2011)
(35791 min = 596.5 hour = 24.9 days)

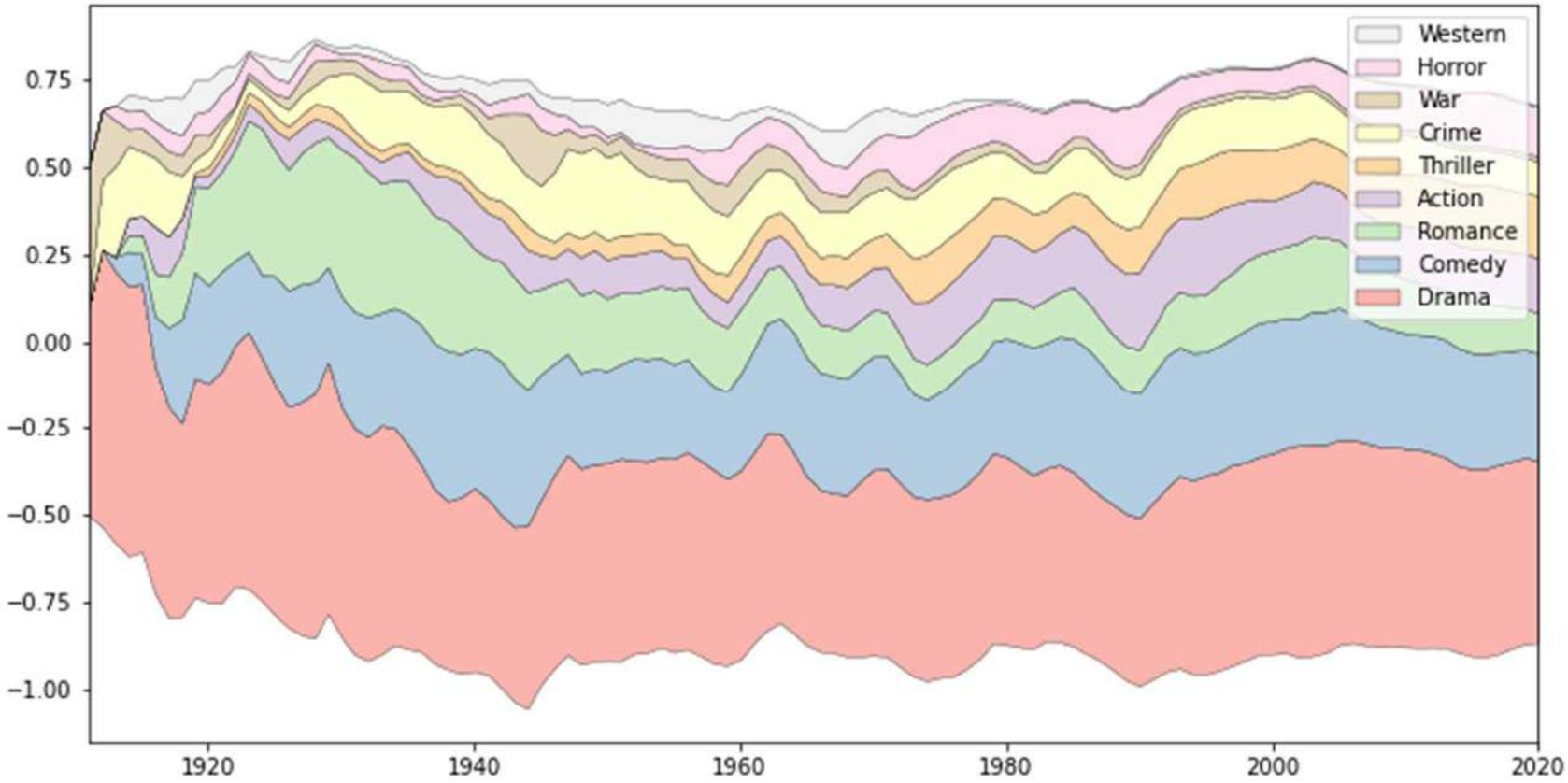
장르 분석 예시

- Kaggle dataset 중 stefanoleone992/imdb-extensive-dataset
  - 영화 85,854편



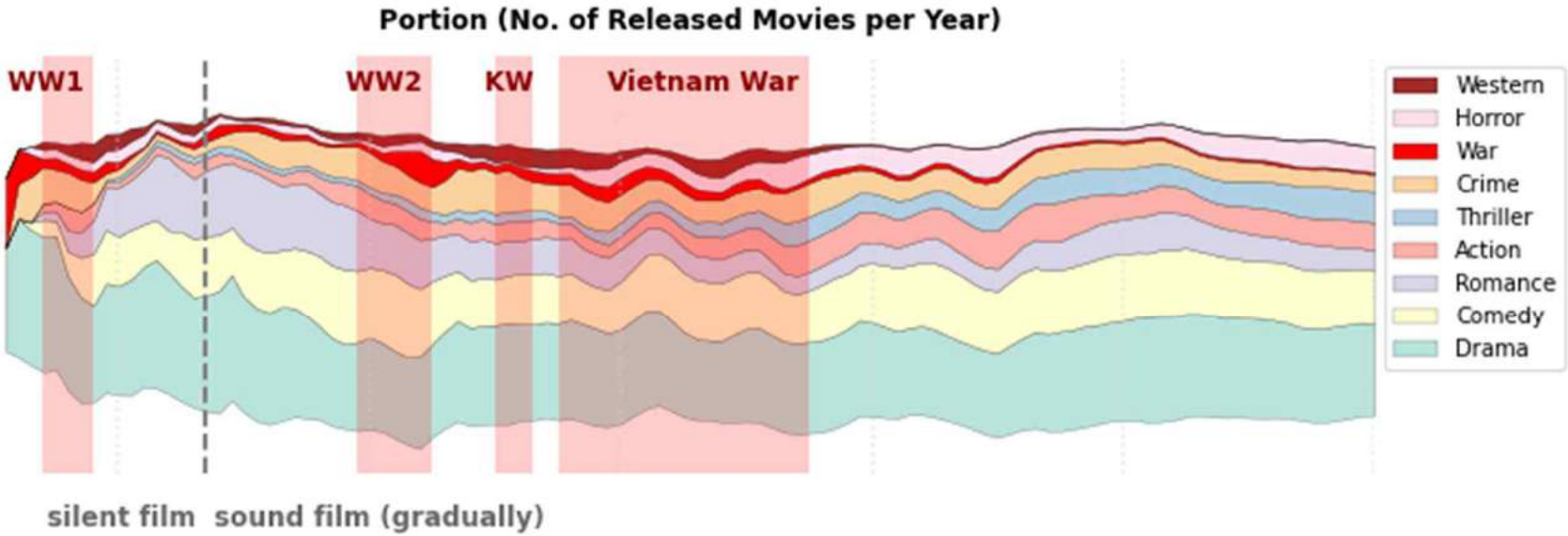
https://bit.ly/3Cy73HB

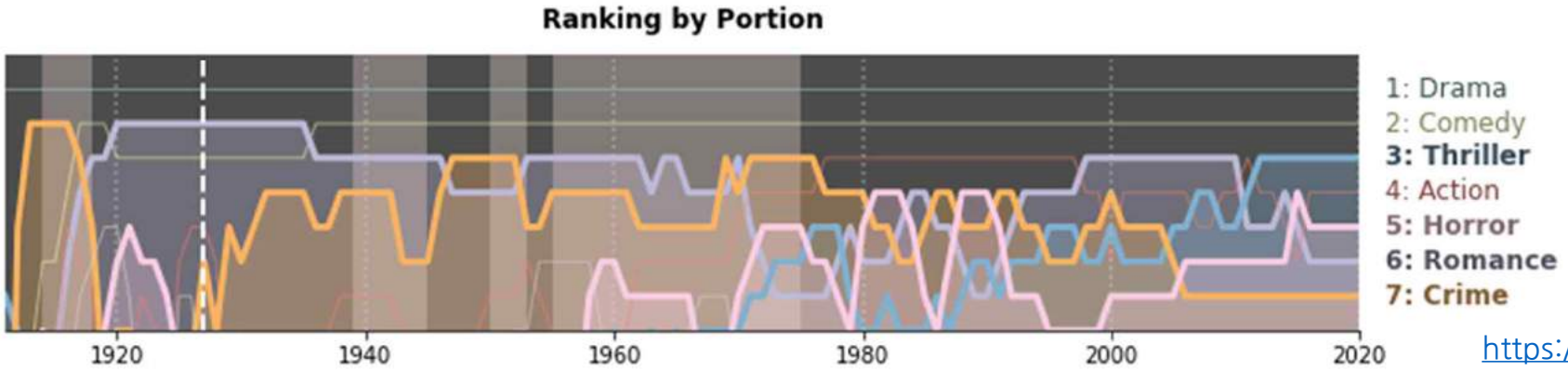
장르 분석 예시 : 장르 경향



https://bit.ly/3Cy73HB

장르 분석 예시 : 전쟁 시기 vs 인기 장르





https://bit.ly/3Cy73HB

데멘토 2021

데이터 분석 공모전 7건, 입상작 69건

'서울시 "여성안심홈세트 지원서비스" 정책 분배적절성, 일회용컵 사용 절감을 위한 그린 뉴딜정책 개선방안, 성공적인 도시 재생 뉴딜사업을 위한 선정지표 개선 연구,
지역별 건강 격차에 따른 의료 자원 조정의 필요성 검토, 서울시 버스 혼잡도 예측 통한 다람쥐버스 신규 노선 제안, 신도시 타당성 요인 분석,
서울시 복지 사각지대 해소를 위한 간이 이동노동자 쉼터 입지선정, 서울시 야생 조류 인공구조물 충돌 사고 우선 조치 지역 제안, MZ세대 소비 트렌드 분석을 통한 서울시 제로페이 활성화 방안,
서울시 스마트폴 구축을 위한 기능 맞춤형 우선 입지 선정, 지하철 공실 문제 해결을 위한 공유 창고 "또타스토리지" 입지 선정, 서울시 열선도로 우선 입지 선정,
서울시 수소차 충전소 우선 입지 선정, 서울특별시 소규모 도심형 물류센터 입지분석, 제주도 관광지 내 효율적인 자동심장충격기(AED) 사용을 위한 입지분석,
인천시민의 야간 골목길 "빅데이터 보안관", 공공데이터 분석을 통한 스마트쉘터 우선지 선정, 데이터 분석으로 금연구역에서의 흡연 예방 디자인 적용,
공공데이터 단계별 시각화를 활용한 주민기피시설 수용성 제고 방안, 부산광역시 출퇴근 혼잡 및 소요 시간 감소를 위한 지하철도 1,2호선 급행 열차도입, 최적 정차역 선정,
시민 공원 이용 만족도 향상 및 효율적 민원 처리를 위한 공원 민원 빅데이터 분석, 신월-신정 다람쥐버스 도임, 디지털 금융 배움터 입지 선정, 수원시 장애인 편의시설 정보 분석,
경기도 청년통장 데이터 분석을 기반으로 한 청년정책 개선방안 제안'

'전통시장 DT 활용 방안, 인공지능을 활용한 가계금융건강검진, 빅데이터를 활용한 사업부지 맞춤형 컨설팅, 지역의 사회구조적 특성이 빈집 형성에 미친 영향 분석,
골목시장 방송 프로그램의 골목시장 활성화 효과 분석, 수출액 예측을 통한 수출 유망 국가와 품목 추천, 코트라 차년도 수출액 예측 과제,
클러스터링을 활용한 무역포트폴리오 다양화, KOTRA 수출 유망국가 추천, 경제적,
산업구조적, 문화적 요인을 기반으로 한 주요 국가의 한국 품목별 수입액 예측모형 개발: 한국의, 한국에 대한 문화적 요인을 중심으로, 분리학습 모델을 통한 국가별 품목별 수입액 예측 및 유망국가 추천,
한국 수입액 예측을 통한 유망 시장 탐색, MZ세대 소비 트렌드 분석을 통한 서울시 제로페이 활성화 방안, 지하철 공실 문제 해결을 위한 공유 창고 "또타스토리지" 입지 선정,
일회용품 쓰레기 감소를 위한 다회용기 렌탈 사업 비즈니스 모델 개발, 서울특별시 소규모 도심형 물류센터 입지분석,
통행량 기반 수도권 최적 택시사업구역 도출, 경기도 청년통장 데이터 분석을 기반으로 한 청년정책 개선방안 제안'

데이터 분석, 의심에서 전달까지: https://jehyunlee.github.io/2021/10/14/Python-DS-86-datadoubt/

데이터 분석: "정책"

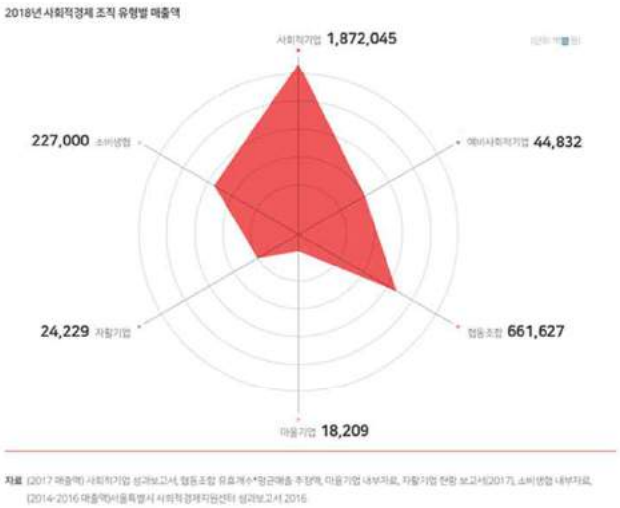
내용 전달: 글 + 그림

설득력: 대안의 장점과 단점

신뢰성: 검증 결과, 예상 오차

① 현황 분석 → ② 대안 제시

예측모델 개발

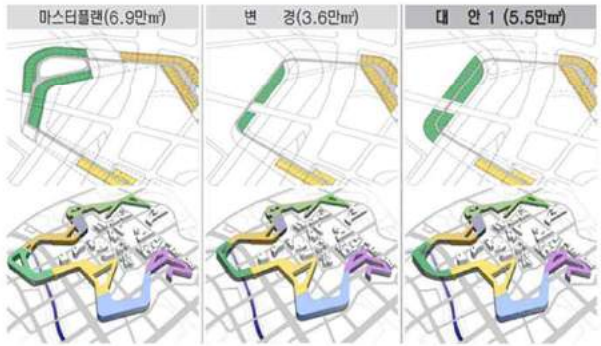


서울시 사회적경제 조직 현황 분석 및 주요 성과 연구 보고서

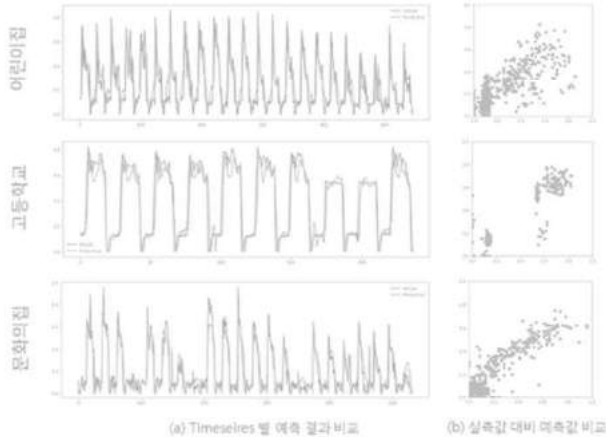
[제 1 안]

▫ 장래 확장가능성 등을 감안하여 기관별 2, 3단계 예비사무실(순사무실의 10%) 중 일정부분(7%) 등을 3-1구역에 집중 배치하여 기관 신설 등에 대응

- **3-1구역 면적** : (당초)6.9만㎡ → (변경)3.6만㎡ → (대안)5.5만㎡
- **장점** : 청사 배치기준 등의 큰 변동이 없어 사업추진 원활하며, 1단계·2단계와 최대한 가까운 배치에 따른 연결통로 단축으로 예산절감이 가능하고 향후 수요 증가에 따른 증축 용이
- **단점** : 마스터플랜 안보다 볼륨이 작아 기본이념 구현에 다소 미흡



정부세종청사 건립백서



건물 열에너지 수요 예측모델

정책 데이터 분석: “어떻게”

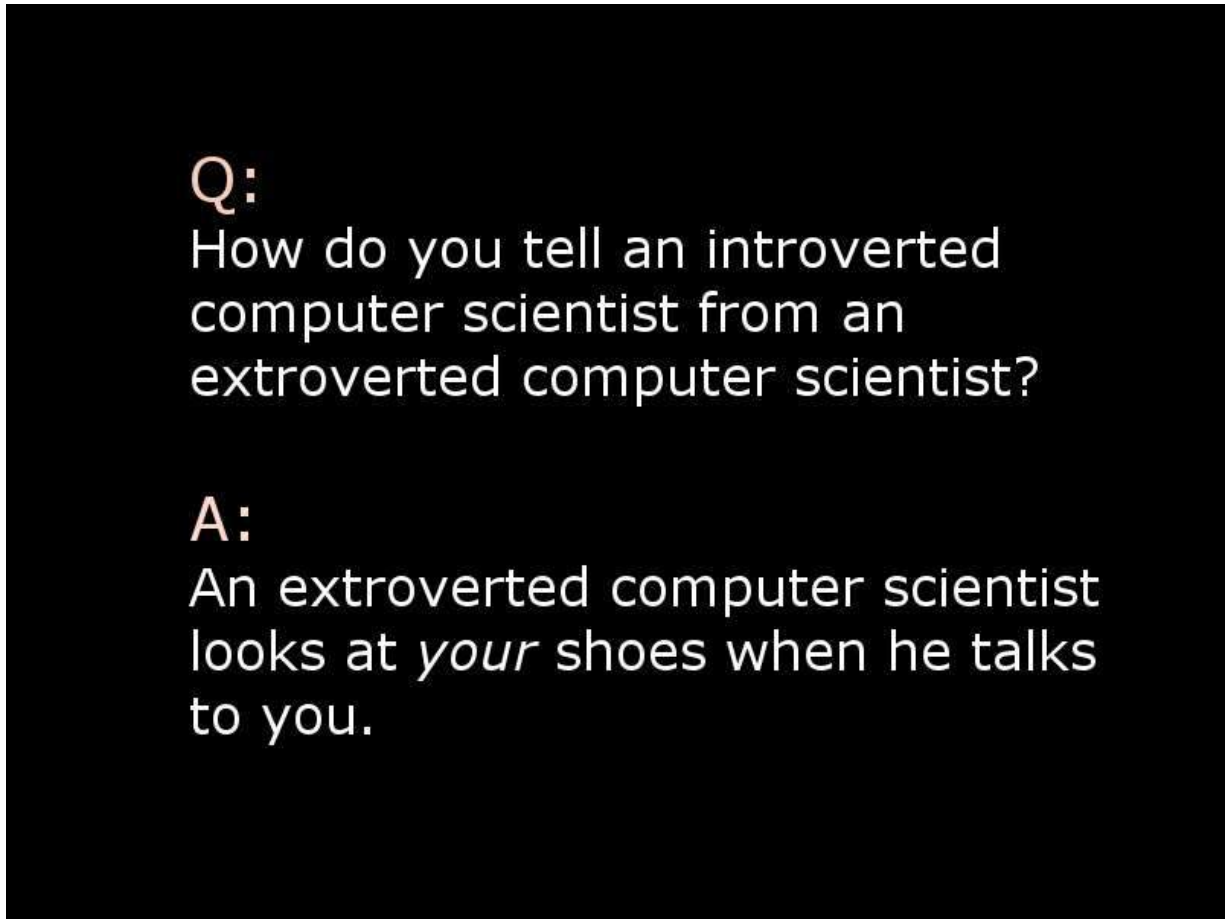
- **장그래:** 사무실도 현장이라는 뜻입니다. 그 현장의 전투화, 당신에게 사무 현장의 전투화를 팔겠습니다.

- **한석율:** 안 사겠습니다. 사무실이 현장이라니 말장난이 지나치시군요. 현장이 뭔 줄이나 아십니까? 사무실 끄적임 몇 번으로 쉽게 쉽게 잘려나가는 구조조정 최하층에서 근무하는 사람들, 현장 노동자라고 합니다! 그들의 전투화를 소개해드릴까요? 워커 신고 일합니다. 무거운 공구가 떨어지면 발등 아작나니까!
  전투화란 그런 겁니다. 전 당신 물건 사지 않겠습니다!

- **장그래:** 한석율 씨는 처음 만났을 때부터 현장을 강조했습니다. 아니, 현장만을 강조했죠. 한석율 씨가 생각하는 현장의 치열함은 기계가 바쁘게 돌아가고 힘을 들여 제품을 만들고 옮기는 것인가 봅니다.
  하지만, 매일 지옥철을 겪으면서 출근하고 제품 수익을 위해 환율과 국제통상 가격을 매일 체크하고, 숫자 하나 때문에 수많은 절차를 두어 실수를 방지하고 문장 하나 때문에 법적 해석을 검토하기도 하고 결과를 집행합니다.
  서류만 넘기면 되는 것이 아닙니다. 밀고 당기는 많은 대화가 있고 그 과정에서 자기 자신이 초라해 보이기까지 하죠. 전화 한 통을 받기 위해 해당국 업무시간까지 밤을 새워 대기하기도 합니다. 한석율 씨가 말하는 현장에서 생산되는 모든 제품은 왜 만들어져야 하는지의 과정을 거친 이후에 존재하는 것입니다. 그 물건들은 사무실을 거치지 않고서는 존재할 수 없는 것입니다.
  제품이 실패하거나 부진을 겪는다는 건 그만큼의 예측 결정에 실패하거나 기획 판단이 실패했다는 겁니다. 공장과 사무는 크게 보아 이어져 있습니다. 제가 생각하는 현장은 한석율 씨가 생각하는 현장과 결코 다르지 않다고 확신합니다.

미생(웹툰, 드라마)

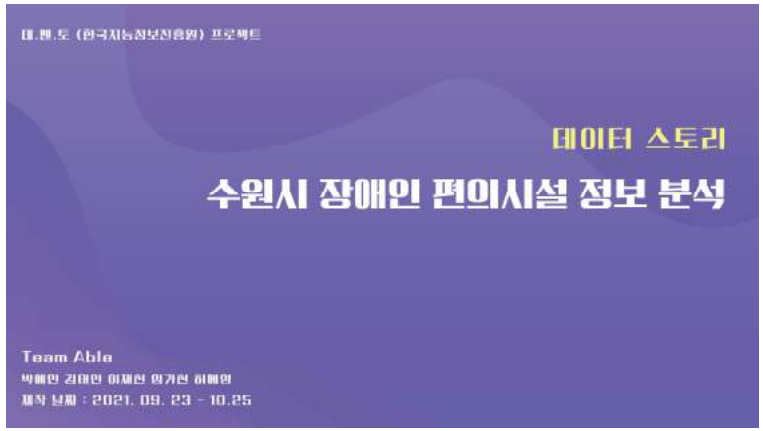
우리들의 MBTI(?)



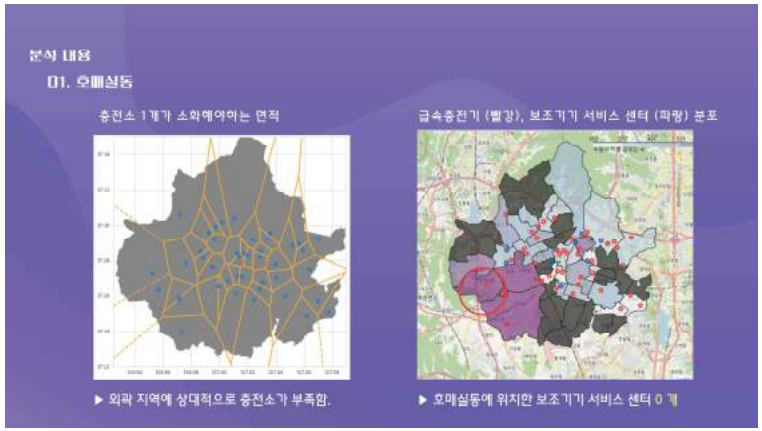
https://www.haeyounglee.com/image/155676898709

데멘토 2021: 수원시 장애인 편의시설 분석

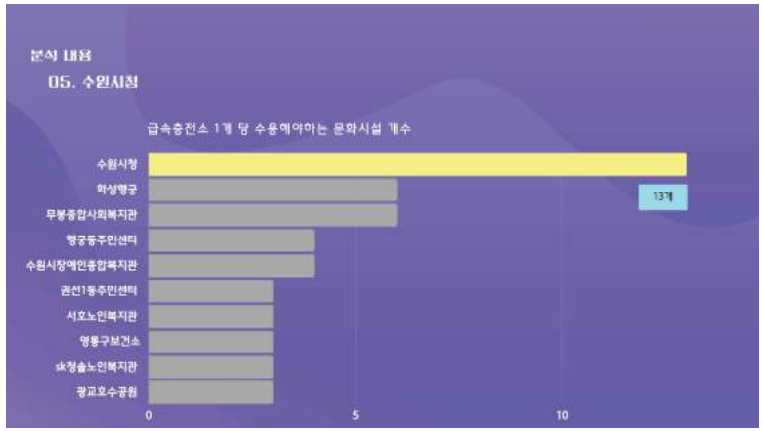
- 최우수상 : Team Able - 박예인, 김태인, 이재헌, 임가현, 허예영



분석



분석



제안

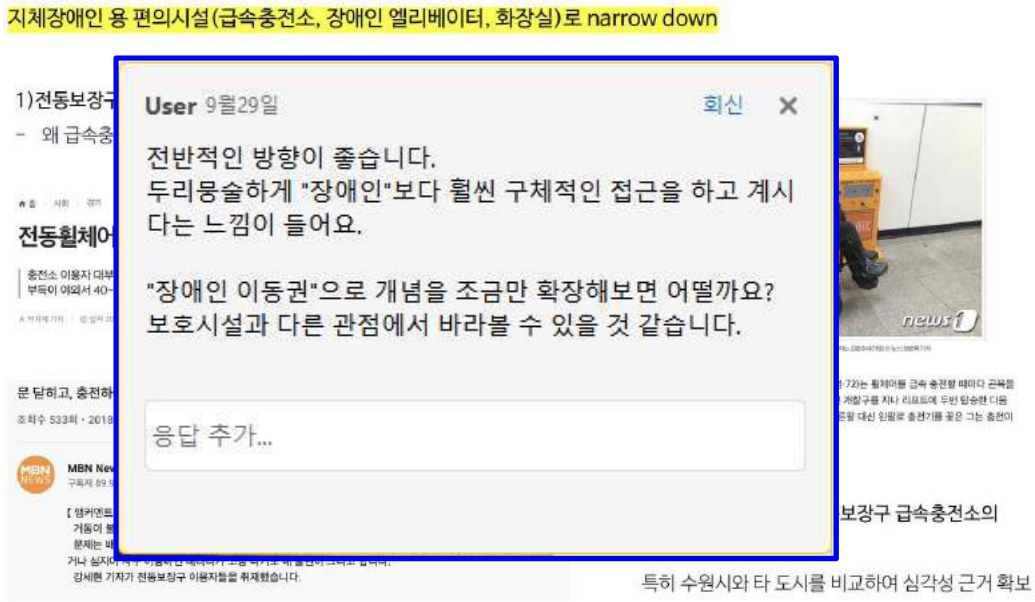
데멘토 2021: 수원시 장애인 편의시설 분석

3. 활동 주제 선정 이유

① 편의시설 장애인 접근성 선정 이유?
- 팀원들의 의견을 모아서 선정
- 롯데마트 안내견 사건 등 시의적절하다고 판단.

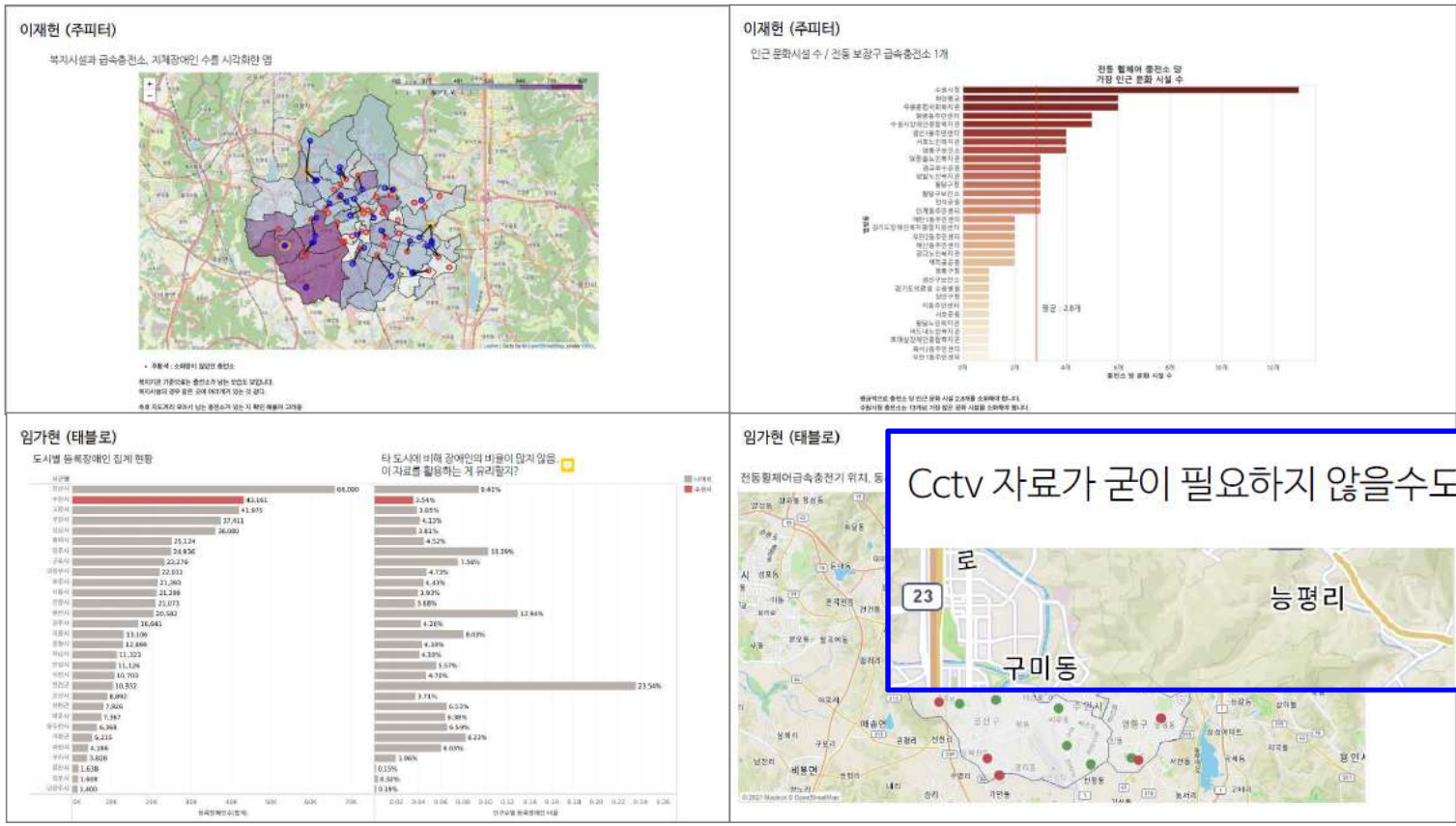
② 수원시 선정 이유?
- 팀원들 중 일부가 살고 있는 도시.
- 경험을 통해 데이터 외적인 면을 포함해 어느 정도 알고 있음.
- 데이터 수집, 분석 결과 타 도시가 더 유리하면 바꿀 수 있음.

③ 현재 진행한 데이터 분석?
- 데이터 현황 확인 및 EDA 진행
- 수집한 데이터를 분석하며 목표 범위 집중: "지체장애인용 편의시설"



데멘토 2021: 수원시 장애인 편의시설 분석

- 데이터 분석 : 최대한 여러 면에서 자세하게



데멘토 2021: 수원시 장애인 편의시설 분석

- 장애 등급 등 규정 파악 + 현장 확인 → 데이터로 알 수 없는 개선안 제시

박예인 - 자폐 장애 vs 지체 장애 연령대별 장애인 수가 차이 나는 이유 :

- 찾지 못했음. 30분 경을 헤매다가 포기함.
- 지하 1층 '만남의 광장'이라는 불명확한 워딩으로 정보를 제공하는 것부터 문제가 있음. 수원역에서 근무하시는 세 분께 여쭤봤을 때도 아무도 정확하게 알려주지 못하심.

2019년 장애인의 다양한 욕구를 반영한 서비스 지원의 기반을 마련하기 위해 장애 등급제가 폐지됨.

박예인

전동보장구 급속충전소 설치 장소 현장 방문

③ 광교 중앙역 환승 센터

▣ 전동보장구 급속충전소 위치: 상행 대합실 중앙

▣ 운영시간 : 24시

▣ 설치 현장 사진 및 설치 실태

- 이외에, 광교 중앙역 장애인 엘리베이터 앞에 놓인 안내문에 시각장애인을 위한 배려 문구가 적혀 있음.
하루에 15,000명의 유동인구를 가진 곳인만큼 이런 사소한 배려들이 필요한 듯함.

박예인

전동보장구 급속충전소 설치 장소 현장 방문

① 수원역

▣ 전동보장구

▣ 운영시간 :

▣ 설치 현장

- 찾지 못

- 지하 1

문제가

정확하

- 경기도

2017년

허예영

전동보장구

① 영통구보건

▣ 전동보장구

▣ 운영시간 :

▣ 설치 현장

- 축시 실외에

보건소 실외

- 영통구 보건

불배 급속충

않음.

[영통구 보건소 옆 영통중앙공원 화장실]

- **차라리 영통구 보건소 바로 옆에 있는 영통중앙공원 화장실 옆(실외)가 설치 장소로 더 적합해 보임.**

▣ **문제점**

- **주말, 공휴일에는 보건소가 문을 열지 않고, 관계자도 없으며, 벨에도 응답하는 사람이 없어 보건소 실내에 있는 전동보장구 급속충전소를 사용하지 못함.**
- **평일이라고 해도 보건소는 9AM-6PM에만 운영하고, 그 시간에도 영통구 보건소가 선별진료소라 코로나 검사를 받으러 오는 사람들이 많아 전동보장구 급속충전소 사용에 어려움이 있을 것으로 보임.**

데멘토 2021: 수원시 장애인 편의시설 분석

- 부록

  데이터 스토리 피드백

  - 전윤선 대표님 (한국접근가능한관광네트워크)

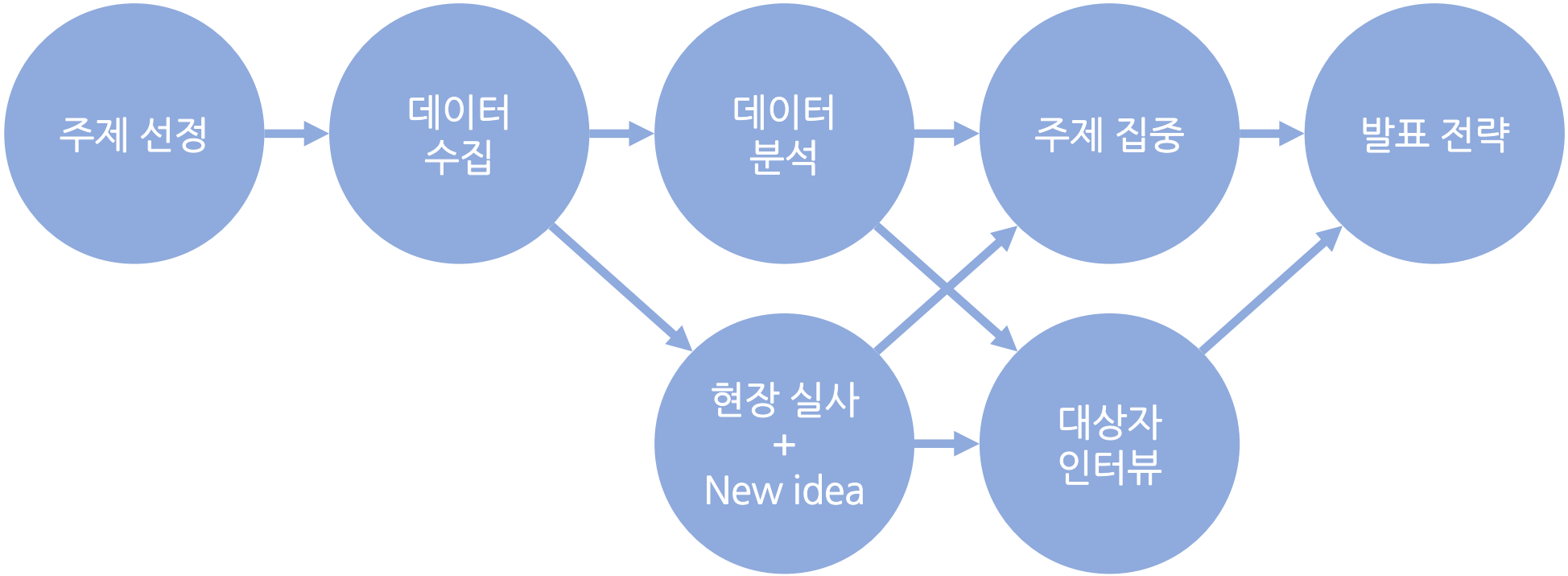
전윤선 대표님

" 장애인을 '보유' 한 행정동이라는 워딩은 맞지 않습니다. 물건이 아니니까요.
장애인이 '가장 많이 거주하는' 행정동 으로 바꾸는 것이 맞습니다."

" 장애인 편의시설 이용층은 장애인에 국한되어있지 않습니다. 어린아이, 노인, 임산부 등을 포함한 모든 신체적 약자를 포함해 만든 법이니, 데이터로 추산된 수치보다 더 많은 사람들이 누릴 수 있는 건강한 대시보드를 만든 것 같아요."

" 전동보장구 급속충전소를 포함한 편의시설의 열악한 실태는, 현재 장애인 관련 모든 복지시설에서 주목하고 있는 시의적절한 주제입니다. 훌륭한 데이터 분석과 현장조사에 감사할 뿐입니다."

데멘토 2021: 수원시 장애인 편의시설 분석



여러분의 멋진 분석을 기대합니다.