03 인기 공유 동영상 데이터 분석 생활 문화

학습 목표
- 문제 해결에 적합한 데이터를 수집하고 분석하여 다양한 학문 분야의 문제를 융합적으로 해결할 수 있다.
- 데이터를 기반으로 자신의 주장을 논리적으로 설명할 수 있다.

1 데이터 기반 문제 해결 과정

데이터에 기반하여 문제를 해결하려면 먼저 문제를 명확하게 정의하고 문제 해결에 적합한 데이터를 수집해야 한다. 수집한 데이터는 사용 목적에 맞게 구분, 관리하여 분석에 적합한 형태로 구조화한다. 다음으로 데이터를 살펴보며 주요 속성을 파악하고, 데이터를 시각화하여 그래프로 표현하고 데이터 간의 관계를 분석한다.

마지막으로 분석 결과를 종합하여 데이터가 지닌 의미를 해석하고, 이를 바탕으로 논리적으로 의견을 주장할 수 있게 된다.

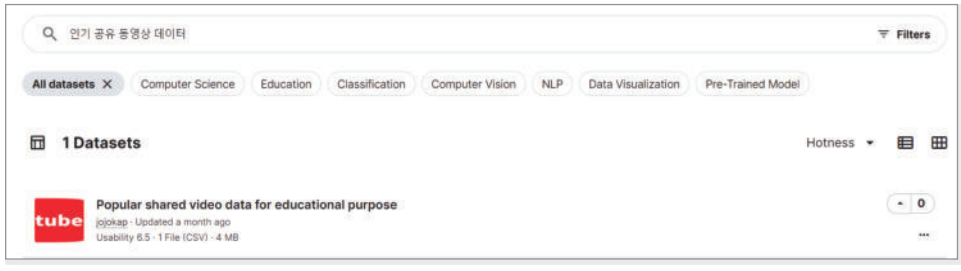
2 문제 해결하기 인기 공유 동영상 데이터 분석

문제 정의 | 영상을 제작하려는 친구에게 동영상 공유 ⊕플랫폼의 데이터를 분석하여 도움을 주려고 한다. 많은 사람이 보고 이용할 수 있는 영상을 만들려면 어떤 것들을 생각해야 하는지 데이터 분석을 통해 알아보자.

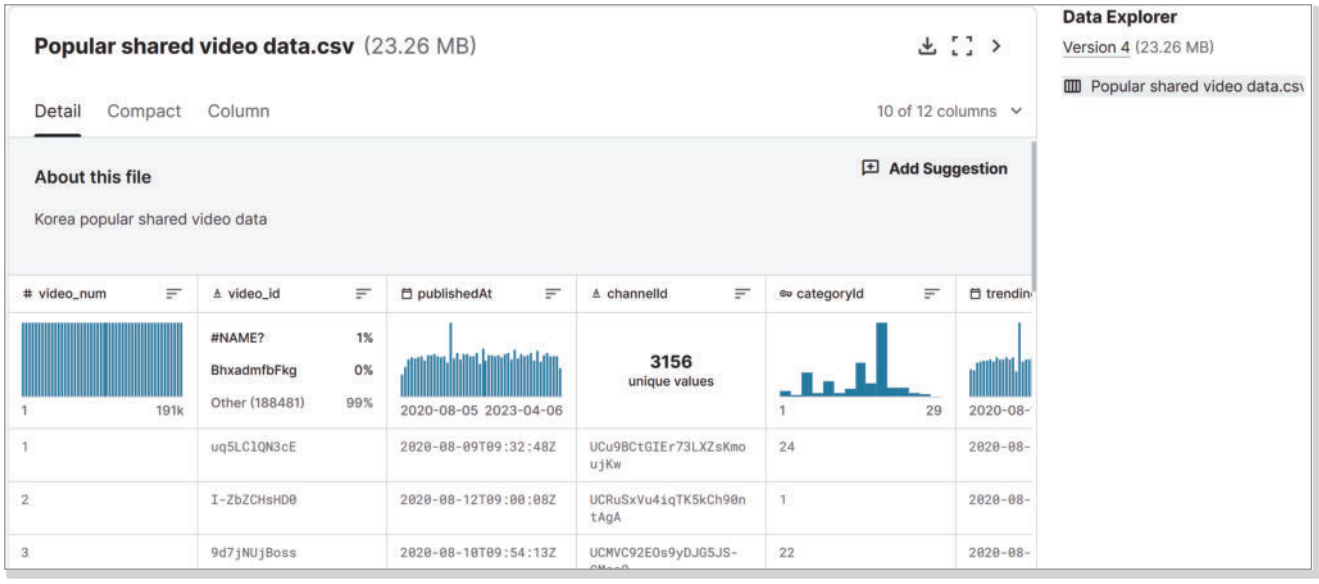
데이터 수집 및 구조화 | 데이터 전문 사이트인 ⊕캐글에 접속하여 화면에 보이는 메뉴 중 [Datasets]을 클릭한다. 검색창에 '인기 공유 동영상 데이터'라고 입력하면 검색 결과에 나오는 교육용으로 편집된 인기 공유 동영상 데이터를 사용한다.

⊕ 플랫폼(platform)
디지털 공간에서 다양한 사람들이 네트워크를 통해 서로 관계를 맺으며 가치를 만들어 내는 서비스 체제를 말한다.

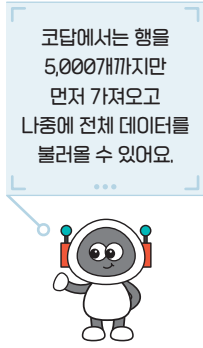
⊕ 캐글(www.kaggle.com)
다양한 데이터를 제공하는 전문 사이트이다. 표로 구조화한 다양한 분야의 데이터를 다운로드할 수 있다.



'Popular shared video data for educational purpose' 데이터를 클릭하면 데이터 소개, 데이터 속성 설명, 데이터 미리 보기, 데이터 다운로드 등의 메뉴가 나타난다. 오른쪽 Data Explorer 부분에서 'Popular shared video data.csv' 파일을 선택하고, 다운로드(⬇)를 클릭하여 데이터를 다운로드한다. 원활하게 데이터를 분석하기 위해 영상 제목, 영상의 설명을 적은 description 속성 등을 삭제한 파일을 이용한다.



코답을 실행한 후 '새 문서'를 클릭하고, 데이터 분석용으로 편집된 'Popular shared video data.csv' 파일을 코답에 드래그하여 표로 구조화된 데이터를 가져온다. 다음은 불러온 데이터의 일부를 나타낸 것이다.



인덱스	video num	video id	publishedAt	channelId	categoryId	trending date	view count	likes	dislikes	comment count	comments disabled	ratings disabled
1	81	6d33R5...	2020-08-10...	UC_pwl...	10	2020-08-12T00:00:...	326895	47622	294	2814	FALSE	FALSE
2	88	nlx-H5k...	2020-08-0...	UCrLQOo...	10	2020-08-12T00:00:...	1815842	31481	1798	2186	FALSE	FALSE
3	95	F1gLID2...	2020-08-0...	UCAkWp...	10	2020-08-12T00:00:...	68512	7260	23	1507	FALSE	FALSE
4	113	FaZ9N...	2020-08-0...	UC9w-h...	24	2020-08-12T00:00:...	803454	12492	125	1707	FALSE	FALSE
5	144	vAmsbv...	2020-08-12...	UCsOW9...	24	2020-08-13T00:00:...	622892	9596	142	1152	FALSE	FALSE
6	183	UKkn5S...	2020-08-12...	UC9kml...	26	2020-08-13T00:00:...	315904	8953	110	1036	FALSE	FALSE
7	187	HOud0...	2020-08-11...	UC6erID...	24	2020-08-13T00:00:...	451460	6818	88	1004	FALSE	FALSE
8	202	Wh6c_4...	2020-08-12...	UCxeWK...	22	2020-08-13T00:00:...	178743	2679	504	727	FALSE	FALSE
9	290	0dsnm...	2020-08-11...	UC0SoP...	23	2020-08-14T00:00:...	566357	27156	167	2762	FALSE	FALSE
10	306	7Y8Vv_...	2020-08-0...	UCIuFnJr...	23	2020-08-14T00:00:...	1539992	23483	1560	1252	FALSE	FALSE
11	335	2ErtcO2...	2020-08-10...	UCg_IS-...	23	2020-08-14T00:00:...	1320472	36071	3167	12993	FALSE	FALSE
12	339	eFxAlvY...	2020-08-12...	UC5xK2...	26	2020-08-14T00:00:...	166906	2994	54	220	FALSE	FALSE
13	384	sF5tpO...	2020-08-0...	UCSngh...	25	2020-08-14T00:00:...	168340	10730	117	1071	FALSE	FALSE
14	421	mbj73Y...	2020-08-13...	UCJ_onu...	25	2020-08-15T00:00:...	106009	2669	45	505	FALSE	FALSE
15	504	5UahGc...	2020-08-0...	UC0SoP...	23	2020-08-15T00:00:...	3715135	189692	1688	21016	FALSE	FALSE
16	625	q4dSqK...	2020-08-13...	UCSrUOj...	24	2020-08-16T00:00:...	830403	143560	310	5032	FALSE	FALSE
17	658	oQHdE...	2020-08-11...	UCLuCG...	20	2020-08-16T00:00:...	238762	1340	27	627	FALSE	FALSE
18	705	z6q98b...	2020-08-13...	UCbD8E...	24	2020-08-17T00:00:...	1701397	42930	1187	3449	FALSE	FALSE

[그림 II-21] 교육용 인기 공유 동영상 데이터(Popular shared video data.csv)

categoryID 숫자의 의미

1 영화/애니메이션
2 자동차/교통수단
10 음악
15 애완동물/동물
17 스포츠
19 여행/이벤트
20 게임
22 사람/블로그
23 코미디
24 연예/오락
25 뉴스/정치
26 방법/스타일
27 교육
28 과학 기술
29 비영리/활동

데이터 살펴보기 | 데이터의 속성을 살펴보면 video num, view count, likes, dislikes 등 동영상 공유 플랫폼의 다양한 속성들이 나와 있다. 다음을 참고하여 해당 속성을 각각 선택하고 '이름 바꾸기'를 이용하여 데이터 속성 이름을 한글로 바꿔 보자.

데이터 속성

- video num: 동영상 고유 번호
- categoryID: 영상 종류(영상이 어떤 종류의 주제에 해당하는지 번호로 표시)
- view count: 조회 수
- likes: 좋아요 수
- dislikes: 싫어요 수
- comment count: 댓글 수
- comments disabled: 댓글 금지 여부
- ratings disabled: 평가(좋아요/싫어요) 금지 여부

데이터 분석에 필요하지 않은 나머지 속성들을 선택하고 '속성 삭제'를 이용하여 삭제하면 다음과 같다.

데이터의 종류별 구분

- 수치형: 조회 수, 좋아요 수, 싫어요 수, 댓글 수
- 범주형: 동영상 고유 번호, 영상 종류, 댓글 금지 여부, 평가 금지 여부

→ 데이터 속성을 클릭하고, '속성의 특성 편집'을 이용하면 [유형]에서 속성을 변경할 수 있다.

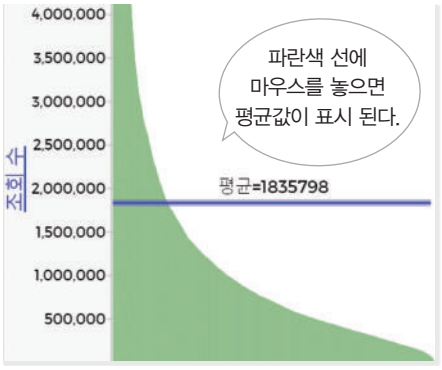
인덱스	동영상 고유 번호	영상 종류	조회 수	좋아요 수	싫어요 수	댓글 수	댓글 금지 여부	평가(좋아요/싫어요) 금지 여부
1	1	24	5947503	53326	105756	139946	FALSE	FALSE
2	14	1	312816	2571	378	236	FALSE	FALSE
3	16	10	372646	24976	228	1124	FALSE	FALSE
4	57	22	534604	0	0	7363	FALSE	TRUE
5	73	24	484236	3225	131	2460	FALSE	FALSE
6	79	25	283867	18522	1781	2416	FALSE	FALSE
7	204	28	8389782	7657	2213	0	TRUE	FALSE
8	209	24	2456311	348686	2240	26933	FALSE	FALSE
9	224	1	651472	5138	275	314	FALSE	FALSE
10	284	24	1755709	45195	397	2265	FALSE	FALSE
11	285	24	165089	6811	77	377	FALSE	FALSE
12	342	23	1069994	24182	2174	5057	FALSE	FALSE
13	462	27	73044	2140	74	158	FALSE	FALSE
14	472	24	128169	7626	21	981	FALSE	FALSE
15	491	10	1505267	141817	1111	5155	FALSE	FALSE
16	503	20	802231	9224	3178	6400	FALSE	FALSE

가장 많은 조회 수를 기록한 영상이 무엇인지 알아보기 위해 '조회 수' 속성을 선택하고 '내림차순 정렬(9→0, Z→A)'을 이용하면 다음과 같이 조회 수 순으로 정렬된 데이터를 확인할 수 있다.

인덱스	동영상 고유 번호	영상 종류	조회 수	좋아요 수	싫어요 수	댓글 수	댓글 금지 여부	평가(좋아요/싫어요) 금지 여부
1	2728	10	262319276	16254784	770144	6303708	FALSE	FALSE
2	66294	10	172293664	12403579	117193	2842820	FALSE	FALSE
3	152891	10	137428027	7941423	0	1183698	FALSE	FALSE
4	40820	10	112816247	7424102	105394	1594992	FALSE	FALSE
5	132950	10	101363270	9614513	0	2588055	FALSE	FALSE
6	136358	10	84959700	5593251	0	360383	FALSE	FALSE
7	39283	10	82295292	6813153	73701	1419486	FALSE	FALSE
8	180504	10	79264364	1736736	0	320815	FALSE	FALSE
9	139348	10	79106132	1568364	0	303135	FALSE	FALSE
10	171103	24	79094667	3781007	0	115461	FALSE	FALSE
11	50628	10	77137095	2402494	54177	344340	FALSE	FALSE
12	50418	10	75246464	2367191	53212	337161	FALSE	FALSE
13	138254	10	74337473	1510847	0	258759	FALSE	FALSE
14	167486	17	65076591	3875747	0	189003	FALSE	FALSE
15	177493	10	64015307	0	0	292039	FALSE	TRUE
16	82802	10	63166762	1882578	88613	209243	FALSE	FALSE

[그림 II-22] 조회 수 순으로 정렬된 데이터

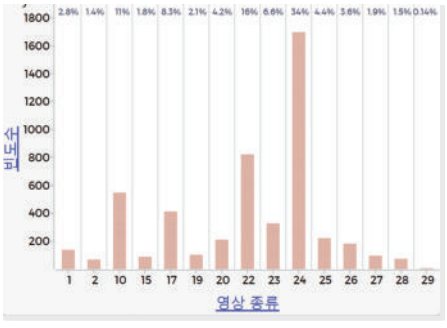
그래프로 표현하기 | 데이터의 조회 수가 어떻게 분포되어 있는지 살펴보기 위해 '그래프'를 이용하면 다음과 같이 나타낼 수 있다. 조회 수가 50만 이하부터 300만 이상인 데이터까지 분포되어 있으며, 평균은 약 183만회임을 알 수 있다.



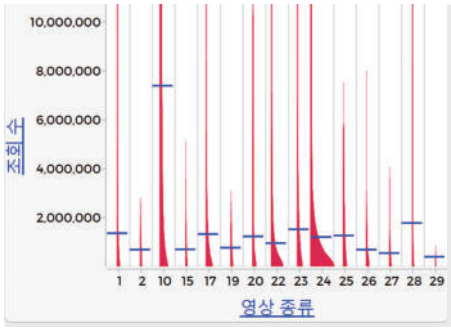
- 세로축을 '조회 수'로 정하고 환경 설정() 메뉴에서 '각 점에 대한 막대'를 선택하여 막대그래프로 나타낸다. 이어서 형식() 메뉴에서 윤곽선을 녹색()으로 선택하여 그래프가 표시되도록 한다.
- 측정() 메뉴에서 '평균'을 클릭하고, 그래프의 분포가 잘 보이도록 조회 수 상한값 숫자를 위쪽으로 적절하게 드래그한다.

❶ 어떤 종류의 영상이 많고, 어떤 종류의 영상이 조회 수가 높은지 ⊕막대그래프를 이용하여 확인하기

'24 연예/오락'이 34 %로 점유율이 가장 높고, 평균 조회 수는 '10 음악'이 가장 많다.



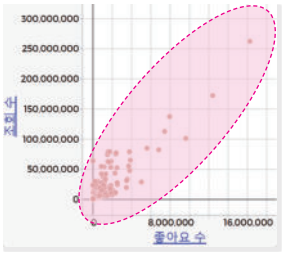
[그림 Ⅱ-23] 영상 종류와 빈도 수 막대그래프



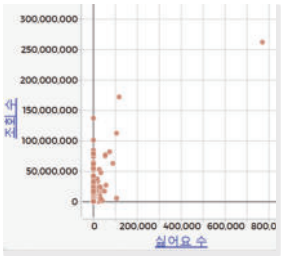
[그림 Ⅱ-24] 영상 종류와 조회 수 평균 막대그래프

❷ 동영상 조회 수와 좋아요 수, 싫어요 수, 댓글 수의 상관관계 확인하기

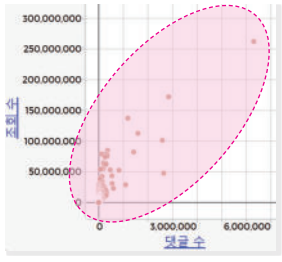
좋아요 수와 댓글 수가 많을수록 조회 수가 증가하는 것으로 보아 조회 수와 좋아요 수, 댓글 수가 관련이 있는 것을 알 수 있다. 그러나 싫어요 수와 조회 수는 대체로 관련이 없는 것으로 나타난다.



▲ 좋아요 수와 조회 수의 산점도 그래프



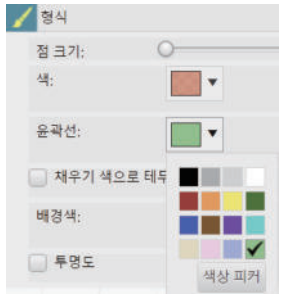
▲ 싫어요 수와 조회 수의 산점도 그래프



▲ 댓글 수와 조회 수의 산점도 그래프

[그림 Ⅱ-25] 동영상 조회 수와 좋아요 수, 싫어요 수, 댓글 수의 상관관계를 나타낸 그래프

[형식] 메뉴에서 윤곽선 설정하기



조회 수 상한값 숫자 조정하기

마우스 드래그

⊕ 막대그래프

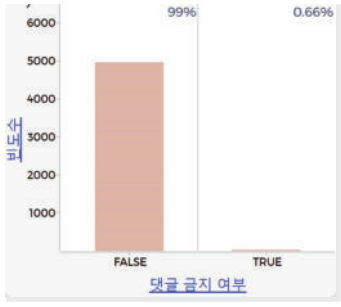
데이터의 양을 막대로 표시하여 나타낸 그래프로 각각의 크기를 쉽게 비교할 수 있다.

링크

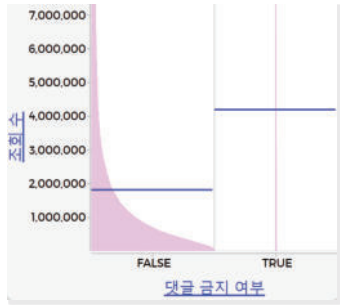
68쪽 산점도 그래프

❸ 댓글 금지 동영상의 비율과 금지 여부, 조회 수에는 어떤 관계가 있는지 확인하기

댓글 금지 여부의 점유율은 FALSE(허용)가 TRUE(금지)보다 많으나 동영상 조회 수의 평균은 오히려 TRUE(금지)가 FALSE(허용)보다 더 높다.



[그림 II-26] 댓글 금지 여부와 빈도 수의 막대그래프



[그림 II-27] 댓글 금지 여부와 조회 수의 평균 막대그래프

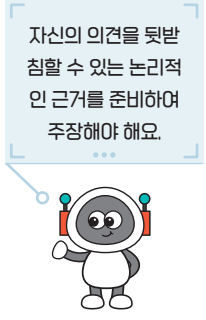
데이터 의미 해석 | 동영상 공유 플랫폼에는 연예/오락, 사람/블로그, 음악, 스포츠, 코미디 관련 영상이 많으며, 평균 조회 수는 음악, 과학 기술, 영화/애니메이션, 스포츠, 게임, 코미디, 뉴스/정치, 연예/오락 순으로 높다.

동영상의 조회 수는 좋아요 수와 댓글 수의 상관관계가 높고, 싫어요 수와는 상관관계가 낮다. 동영상의 댓글 허용 비율이 댓글 금지 비율보다 훨씬 높았으나 댓글을 금지한 영상의 평균 조회 수가 더 높았다는 점에서 다른 중요한 요인이 있다고 생각할 수 있다.

예 데이터 분석을 통해 친구에게 해 줄 수 있는 조언

- 조회 수가 높은 영상을 만들려면 음악, 과학 기술, 영화/애니메이션, 스포츠, 게임, 코미디 등과 같이 특정 전문 분야 영상을 제작하는 것이 좋다.
- 조회 수가 높은 영상을 만들려면 좋아요 수와 댓글 수를 높일 수 있도록 하는 방법을 고민해야 하며, 싫어요와 같은 부정적인 반응은 조회 수와는 관련이 없으므로 민감하게 반응하지 않아도 된다.
- 댓글 금지는 허용하는 것이 그렇지 않은 것보다 비율이 높으니 허용하는 방향으로 한다. 하지만 평균 조회 수를 보면 더 중요한 것은 동영상의 품질이라고 생각할 수 있다.

융합적 문제 해결에 적용하기 데이터 분석 결과와 그에 기반한 주장을 다음과 같이 다양한 분야에서 적용할 수 있다.



예 데이터 분석 결과에 기반한 주장을 융합적인 문제 해결에 적용하기

- 진로 적성 프로젝트에 적용하여 자신의 흥미와 적성에 맞는 진로(전문) 분야를 데이터에 기반하여 찾아보기
- 환경 문제 미술 작품 만들기 프로젝트에 적용하여 적극적인 피드백 데이터로 학생 참여와 환경 문제에 대한 관심 높이기