프로젝트 결과 보고서
코드 실행 결과 별첨

데이터 전처리

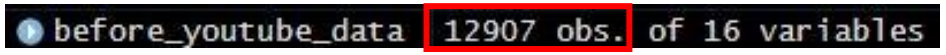
중복 동영상 삭제

```
duplicated(youtube_data$video_id) #데이터 중복 검사

duplicated_youtube_data <- youtube_data[!duplicated(youtube_data$video_id),]
#video_id를 기준으로 중복된 데이터 삭제
View(duplicated_youtube_data)

duplicated(duplicated_youtube_data$video_id) #중복 제거 이후 데이터 중복검사
```

전처리 전



전처리 후

데이터 전처리

새로운 컬럼 생성

category_name	likes_dislikes	tags_count	title_length	description_length	publishedAt_format	trending_date_forma
Entertainment	1.983197690	2	11	0	2020-08-09 09:32:48	2020-08-12
Film & Animation	0.017490440	6	18	496	2020-08-12 09:00:08	2020-08-12
People & Blogs	3.833203516	27	15	0	2020-08-10 09:54:13	2020-08-12
Entertainment	0.048735532	12	20	35	2020-08-11 15:00:58	2020-08-12
Music	0.027630845	13	48	524	2020-08-11 09:00:13	2020-08-12
Music	0.021245945	47	25	301	2020-08-11 15:00:13	2020-08-12
Film & Animation	0.059035602	0	36	23	2020-08-10 09:37:33	2020-08-12
Entertainment	1.022776074	0	19	0	2020-08-11 14:00:01	2020-08-12
Comedy	0.025672186	67	29	810	2020-08-12 09:00:02	2020-08-12
Comedy	0.009258427	10	37	175	2020-08-11 09:30:00	2020-08-12
Education	0.008176265	12	87	896	2020-08-10 09:00:12	2020-08-12
Comedy	0.060512204	50	88	775	2020-08-09 11:30:01	2020-08-12
Entertainment	0.039918573	13	27	55	2020-08-09 16:12:59	2020-08-12
Film & Animation	0.147024504	10	38	148	2020-08-11 07:00:04	2020-08-12
News & Politics	0.014430203	10	61	778	2020-08-12 01:54:26	2020-08-12

새로운 컬럼 생성 실행 결과

영상의 조회수가 높을수록 인기 동영상에 등재될 확률이 높은가?

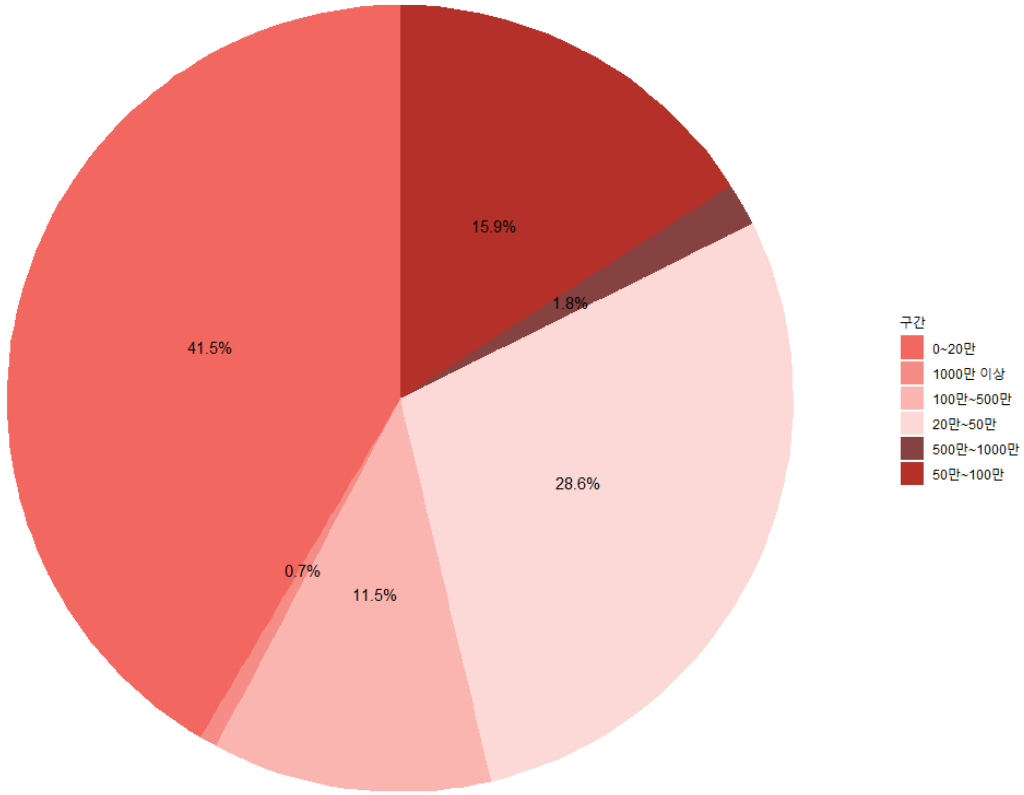
가설 1) 영상의 조회수가 높을수록 인기 동영상에 등재 될 확률이 높을 것이다. 조회수는 최소 100만회 이상이 많을 것 이다.

```
> mean_view_count <- summary(youtube_data$view_count)
> mean_view_count
    Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
   14209   111681   264485   777576   620440 76805026
```

조회수의 평균치를 조회한 결과 약 77만으로 출력

영상의 조회수가 높을수록 인기 동영상에 등재될 확률이 높은가?

가설 1) 영상의 조회수가 높을수록 인기 동영상에 등재 될 확률이 높을 것이다. 조회수는 최소 100만회 이상이 많을 것 이다.



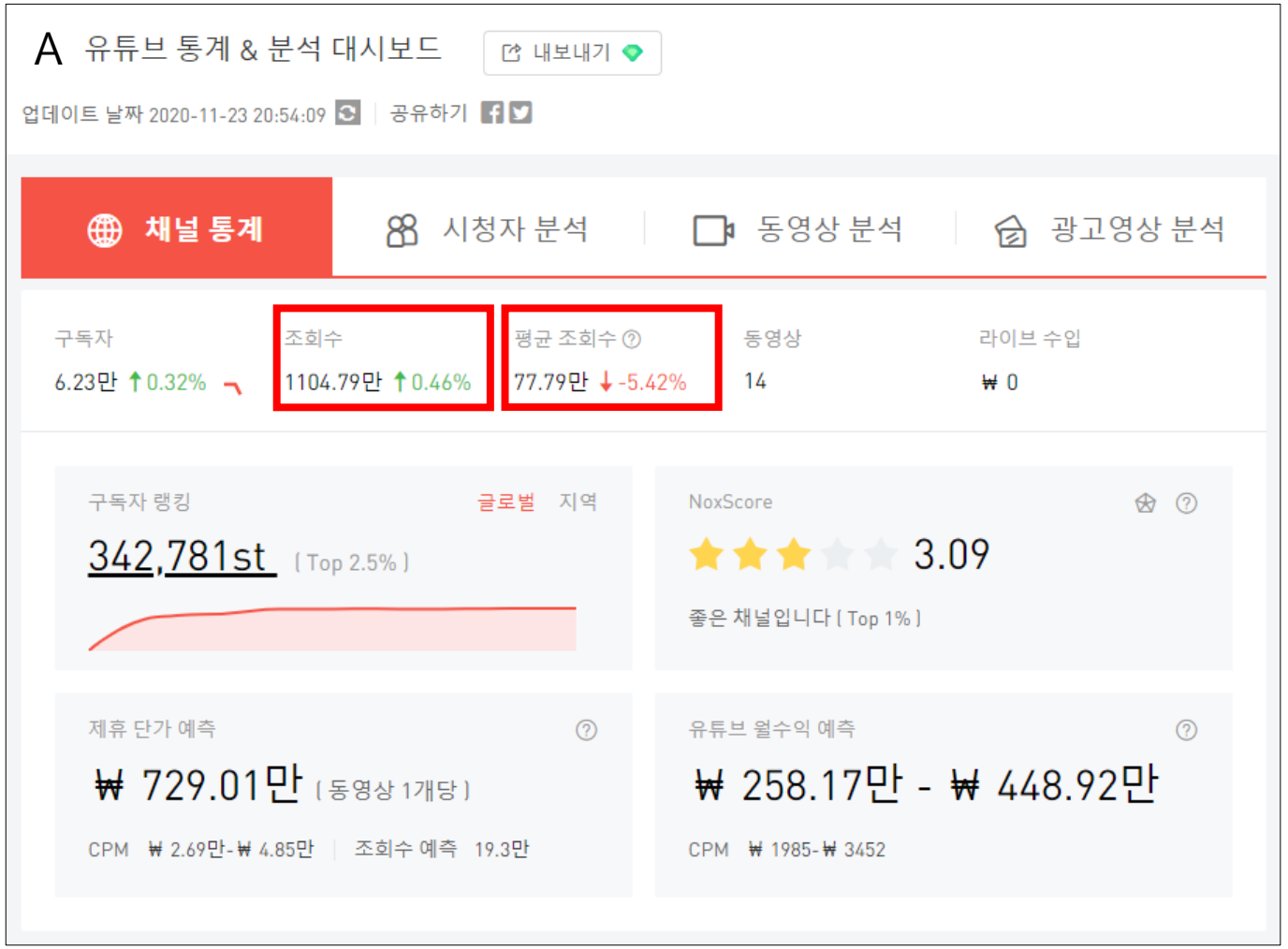
조회수 파이그래프 시각화 결과

50만 이하의 영상 비율 전체의 70.1%

영상의 어떤 요소로 인해 인기동영상에 등재되는가?

변경된 가설 1) 현재 이슈가 되고 있는 것들을 다루는 영상. 즉, 영상의 화제성이 가장 큰 요소로 작용 할 것 이다.

익명



https://kr.noxinfluencer.com/youtube/channel/UC-60AlFBYJEoItr-u-tQXbw

영상의 어떤 요소로 인해 인기동영상에 등재되는가?

변경된 가설 1) 현재 이슈가 되고 있는 것들을 다루는 영상. 즉, 영상의 화제성이 가장 큰 요소로 작용 할 것 이다.

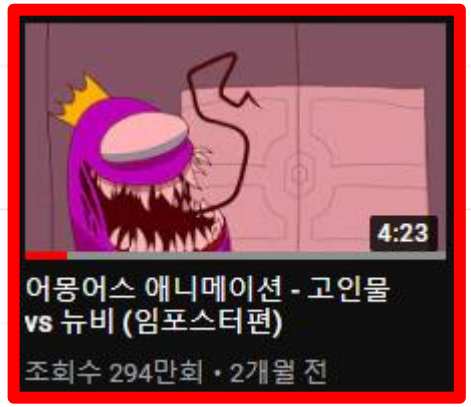
https://datalab.naver.com/keyword/trendResult.naver?hashKey=N_f0975b16ee9f42746be0af4aa5b88af4

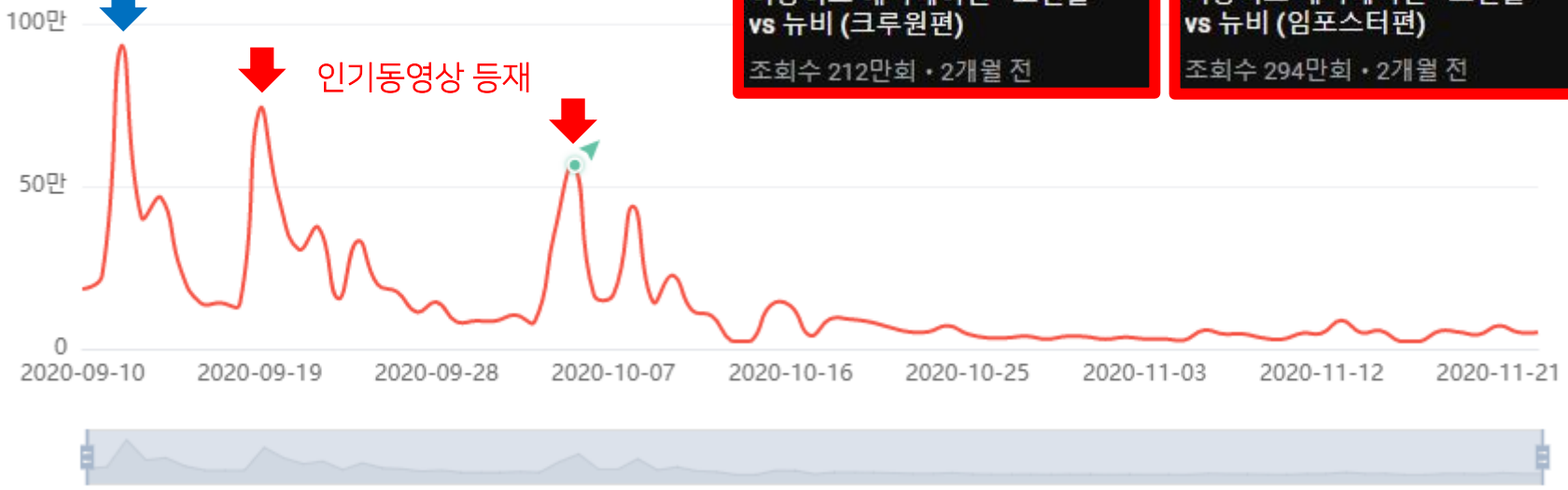
영상의 어떤 요소로 인해 인기동영상에 등재되는가?

변경된 가설 1) 현재 이슈가 되고 있는 것들을 다루는 영상. 즉, 영상의 화제성이 가장 큰 요소로 작용 할 것 이다.

어몽어스 유행
조회수 폭발

인기동영상 등재

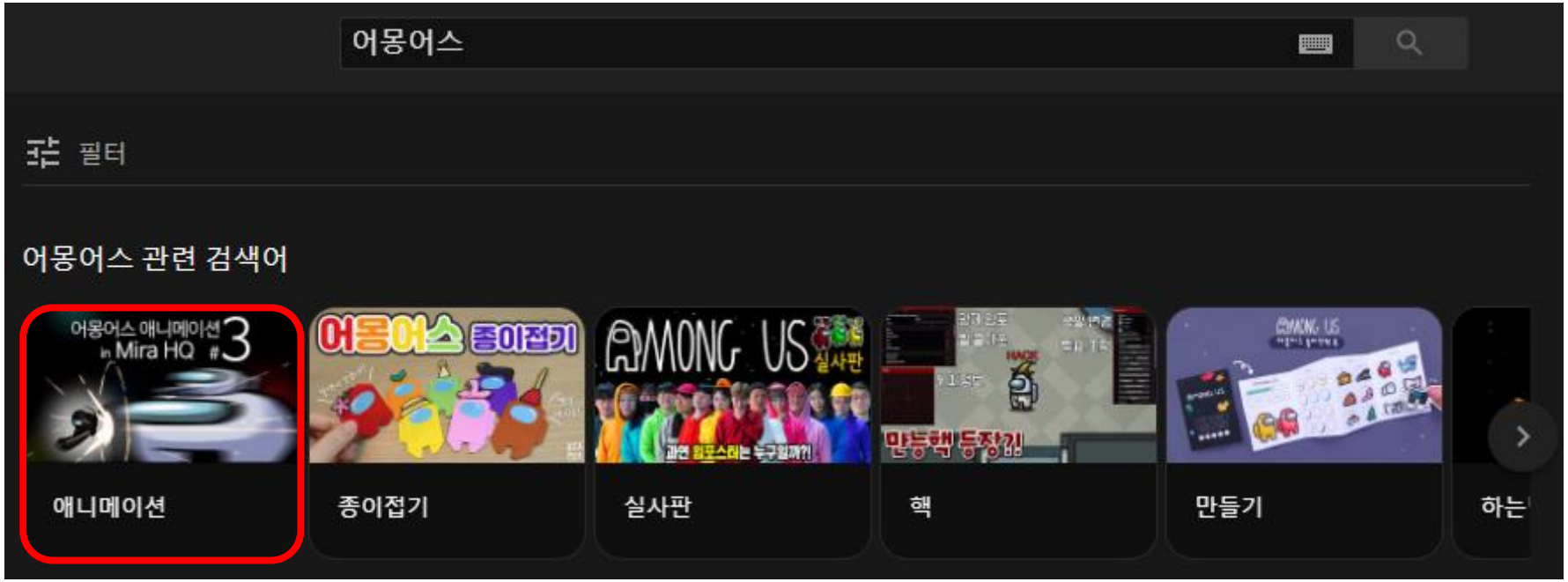




https://kr.noxinfluencer.com/youtube/channel/UC-60AlFBYJEoltr-u-tQXbw

영상의 어떤 요소로 인해 인기동영상에 등재되는가?

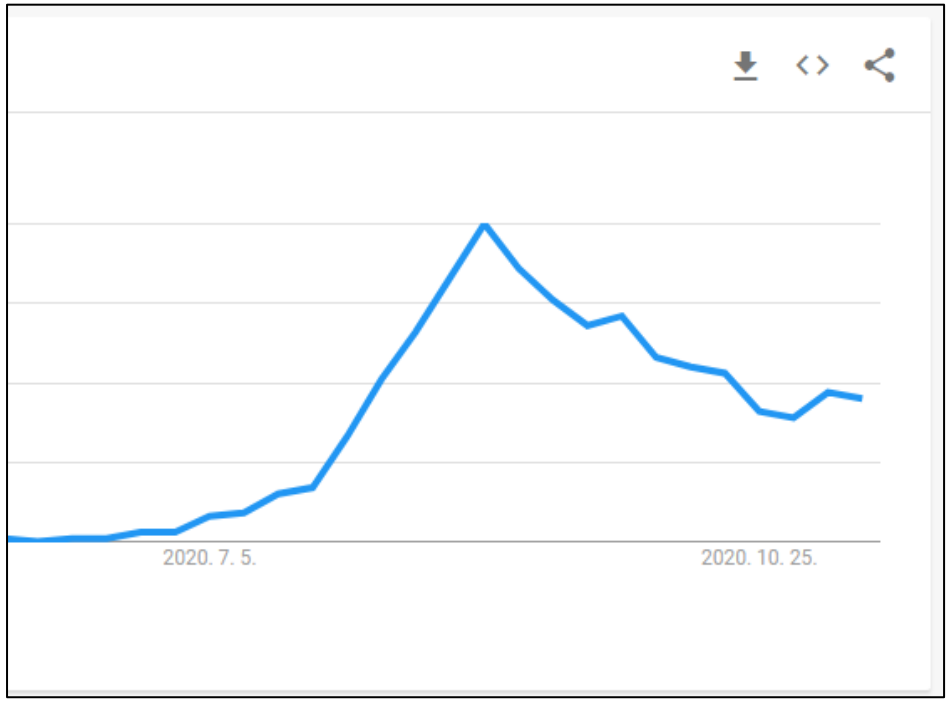
변경된 가설 1) 현재 이슈가 되고 있는 것들을 다루는 영상. 즉, 영상의 화제성이 가장 큰 요소로 작용 할 것 이다.



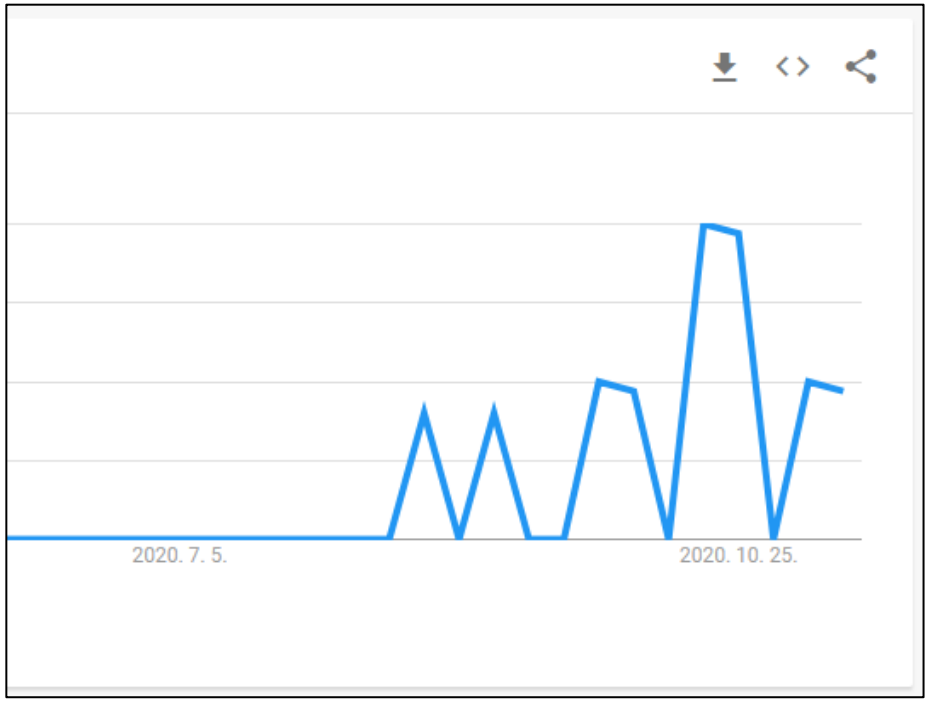
영상의 어떤 요소로 인해 인기동영상에 등재되는가?

변경된 가설 1) 현재 이슈가 되고 있는 것들을 다루는 영상. 즉, 영상의 화제성이 가장 큰 요소로 작용 할 것 이다.

google [어몽 어스] 검색 량



google [어몽어스 애니메이션] 검색 량



영상의 어떤 요소로 인해 인기동영상에 등재되는가?

변경된 가설 1) 현재 이슈가 되고 있는 것들을 다루는 영상. 즉, 영상의 화제성이 가장 큰 요소로 작용 할 것 이다.

https://kr.noxinfluencer.com/youtube/channel/UC-60AlFBYJEoltr-u-tQXbw

영상 업로드 시점 이후 인기 동영상에 등재되기까지의 시간은 얼마나 걸리는가?

가설 2) 인기 동영상 목록이 업데이트되는 최소 주기인 15분 부터 최대 7일까지의 기간으로 예상한다.

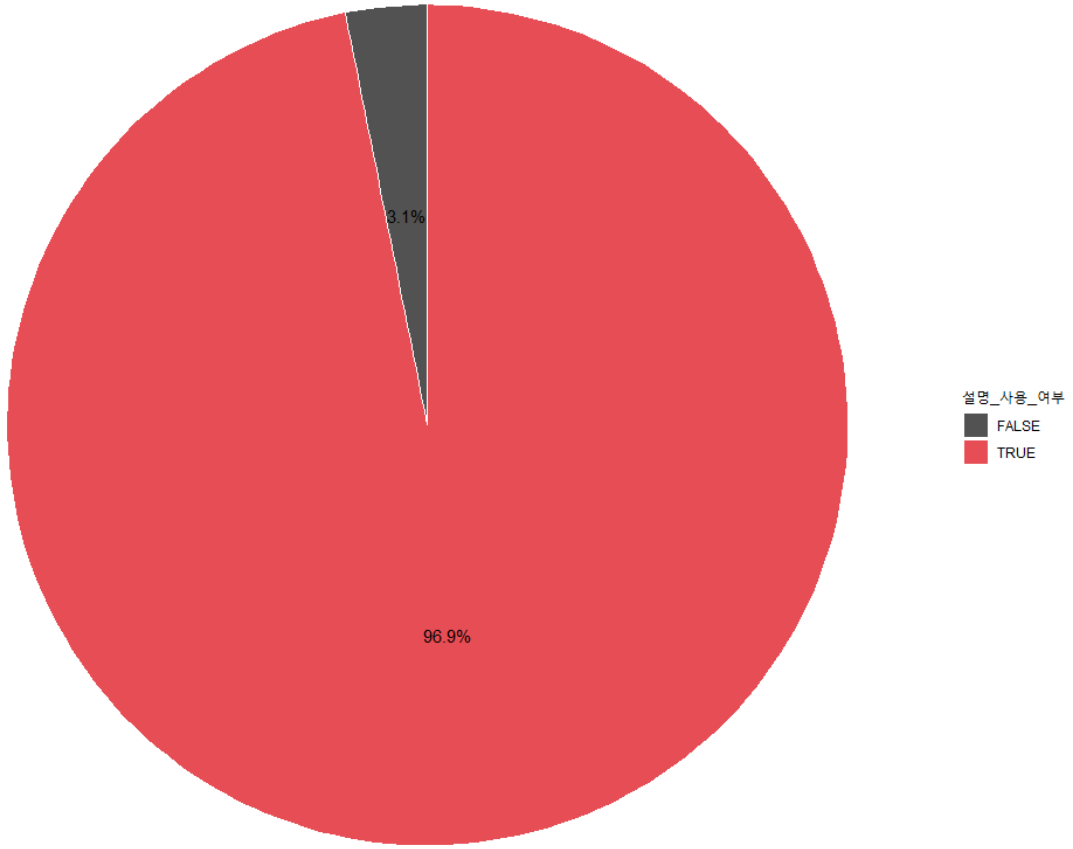
```
> mean_time_cal2 <- mean(mean_time2)
> mean_time_cal2
Time difference of 13.50112 hours
```

```
> max_time_cal3 <- max(mean_time2)
> max_time_cal3
Time difference of 237 hours
```

평균 13.5시간, 최댓값으로 237시간

영상 설명, 태그의 기재는 인기동영상 등재에 영향을 미치는가?

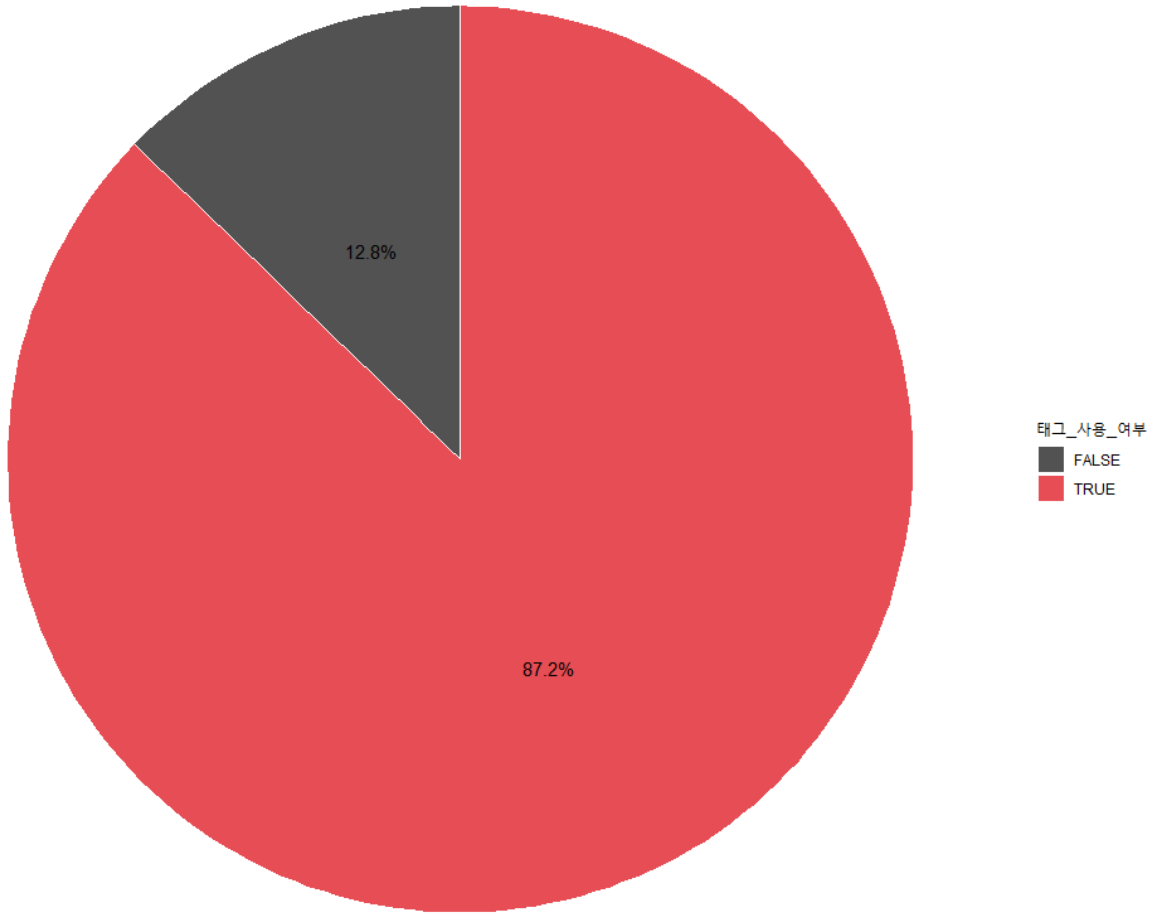
가설 3) 키워드 검색으로 인한 시청자 유입이 쉬워지므로 영상 설명 기재는 인기동영상 등재에 영향을 미칠 것 이다.



영상 설명 사용 여부 비율 시각화

영상 설명, 태그의 기재는 인기동영상 등재에 영향을 미치는가?

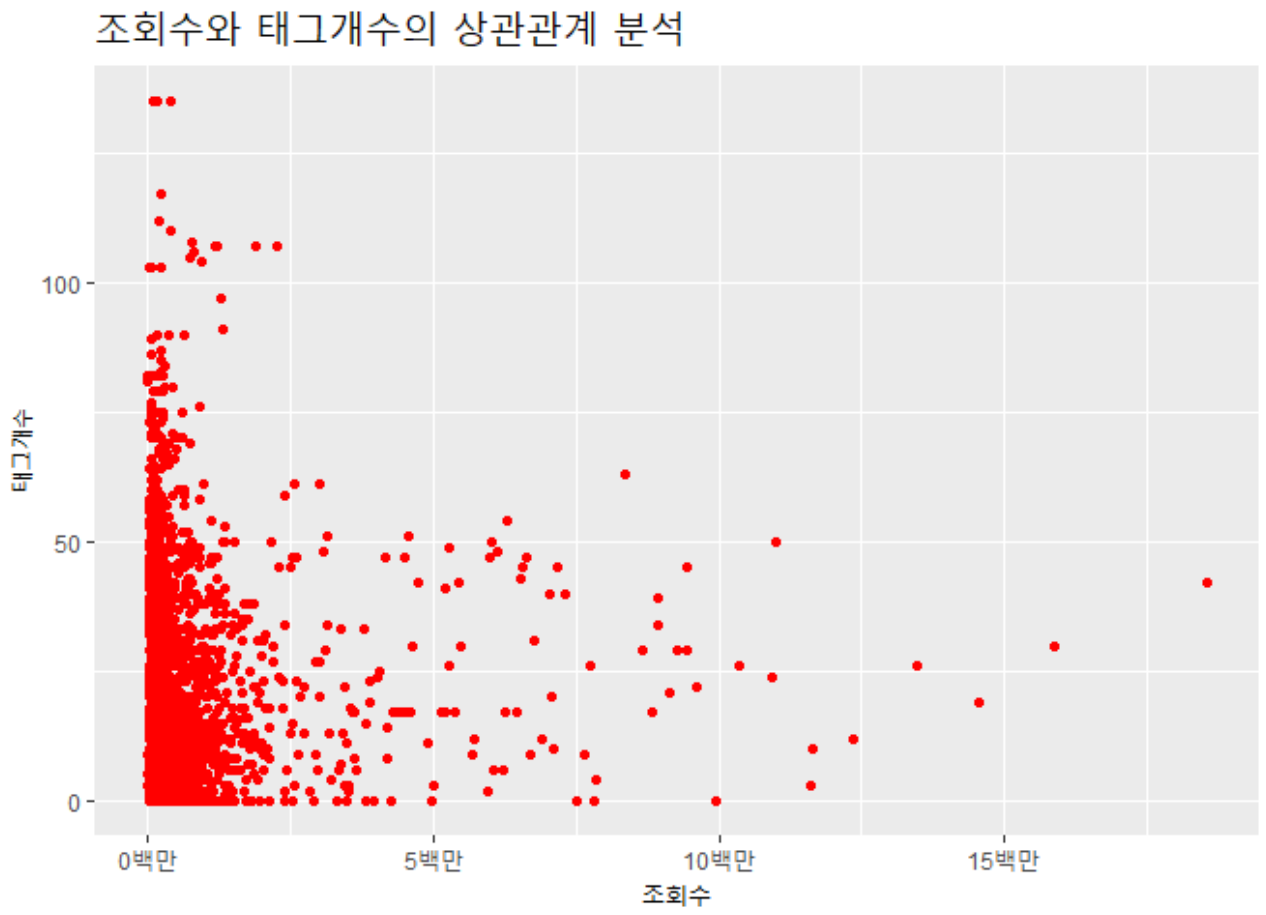
가설 3) 키워드 검색으로 인한 시청자 유입이 쉬워지므로 영상 설명 기재는 인기동영상 등재에 영향을 미칠 것 이다.



태그 사용 여부 비율 시각화

영상 설명, 태그의 기재는 인기동영상 등재에 영향을 미치는가?

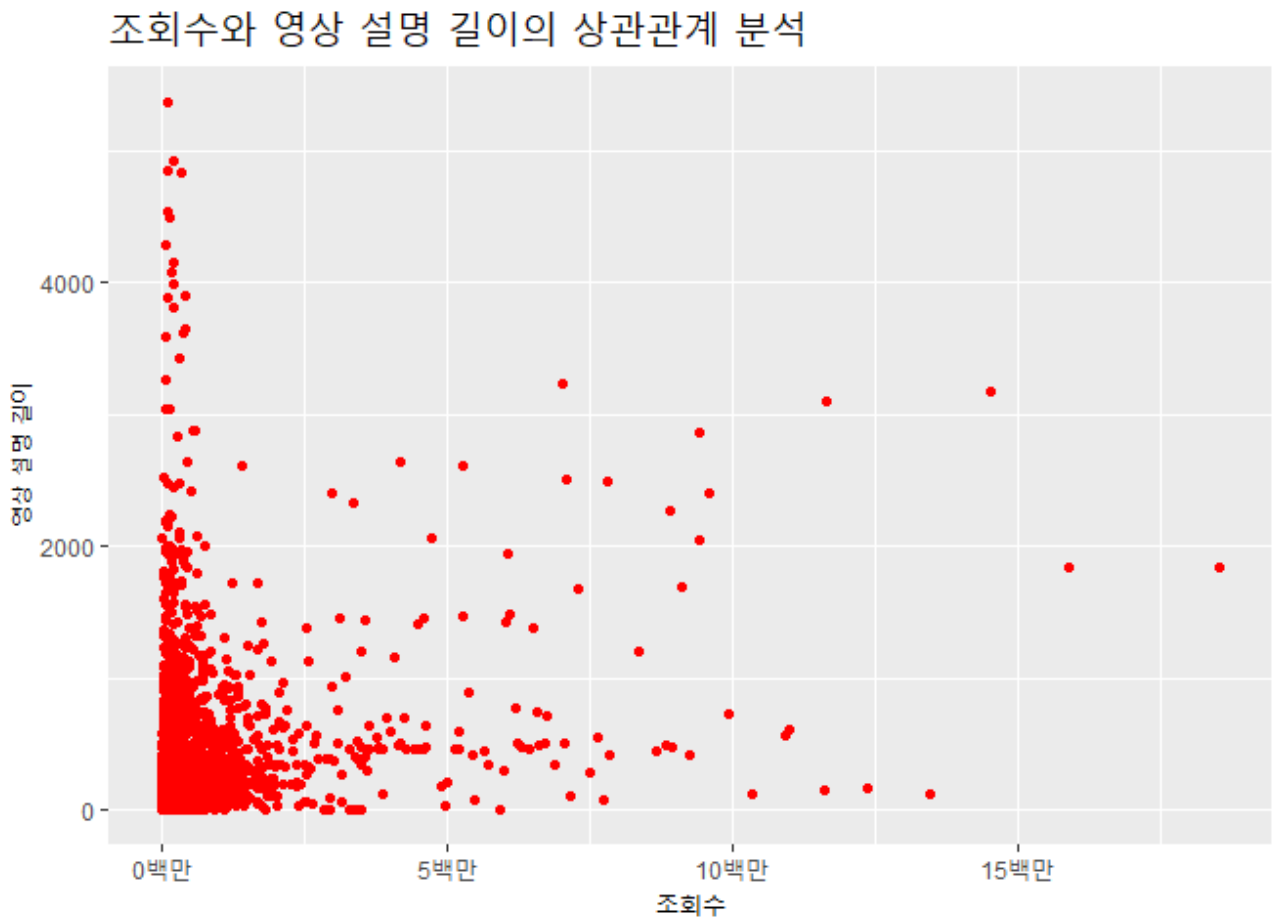
가설 3) 키워드 검색으로 인한 시청자 유입이 쉬워지므로 영상 설명 기재는 인기동영상 등재에 영향을 미칠 것 이다.



조회수, 태그 개수 간의 상관관계 산점도

영상 설명, 태그의 기재는 인기동영상 등재에 영향을 미치는가?

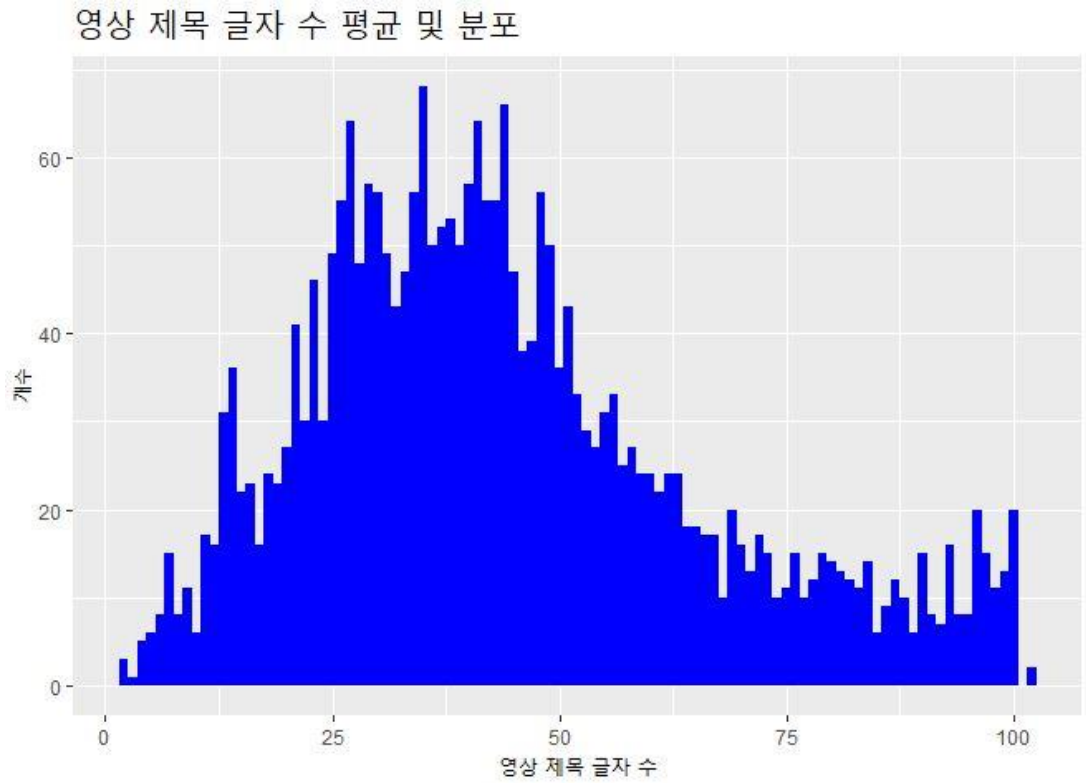
가설 3) 키워드 검색으로 인한 시청자 유입이 쉬워지므로 영상 설명 기재는 인기동영상 등재에 영향을 미칠 것 이다.



조회수, 영상 설명 길이간의 상관관계 산점도

영상 설명, 태그의 기재는 인기동영상 등재에 영향을 미치는가?

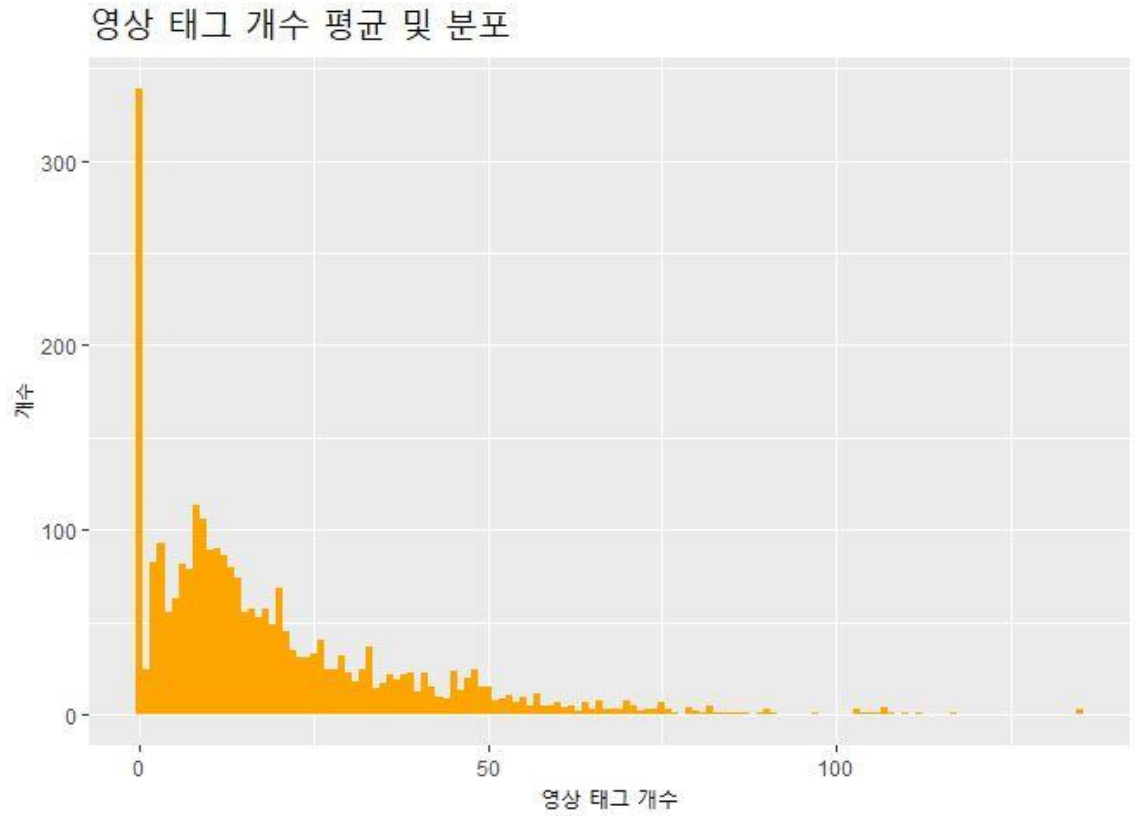
가설 3) 키워드 검색으로 인한 시청자 유입이 쉬워지므로 영상 설명 기재는 인기동영상 등재에 영향을 미칠 것 이다.



영상 제목 글자수 평균 및 분포 히스토그램

영상 설명, 태그의 기재는 인기동영상 등재에 영향을 미치는가?

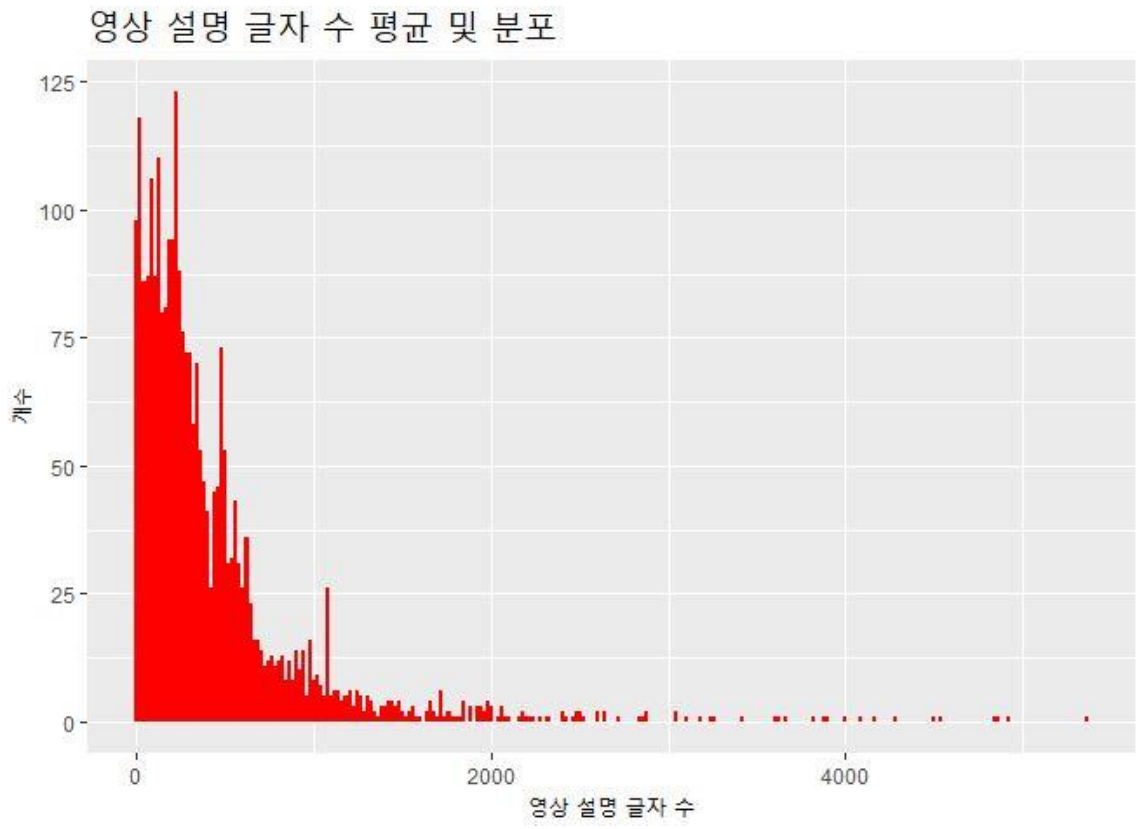
가설 3) 키워드 검색으로 인한 시청자 유입이 쉬워지므로 영상 설명 기재는 인기동영상 등재에 영향을 미칠 것 이다.



영상 태그 개수 평균 및 분포 히스토그램

영상 설명, 태그의 기재는 인기동영상 등재에 영향을 미치는가?

가설 3) 키워드 검색으로 인한 시청자 유입이 쉬워지므로 영상 설명 기재는 인기동영상 등재에 영향을 미칠 것 이다.



영상 설명 글자수 평균 및 분포 히스토그램

인기 동영상의 제목/태그/설명에 자주 사용되는 키워드는 무엇인가?

가설 4) 키워드 검색으로 인한 시청자 유입이 쉬워지므로 영상 설명 기재는 인기동영상 등재에 영향을 미칠 것 이다.

제목 워드클라우드

인기 동영상의 제목/태그/설명에 자주 사용되는 키워드는 무엇인가?

가설 4) 키워드 검색으로 인한 시청자 유입이 쉬워지므로 영상 설명 기재는 인기동영상 등재에 영향을 미칠 것 이다.

태그 워드클라우드

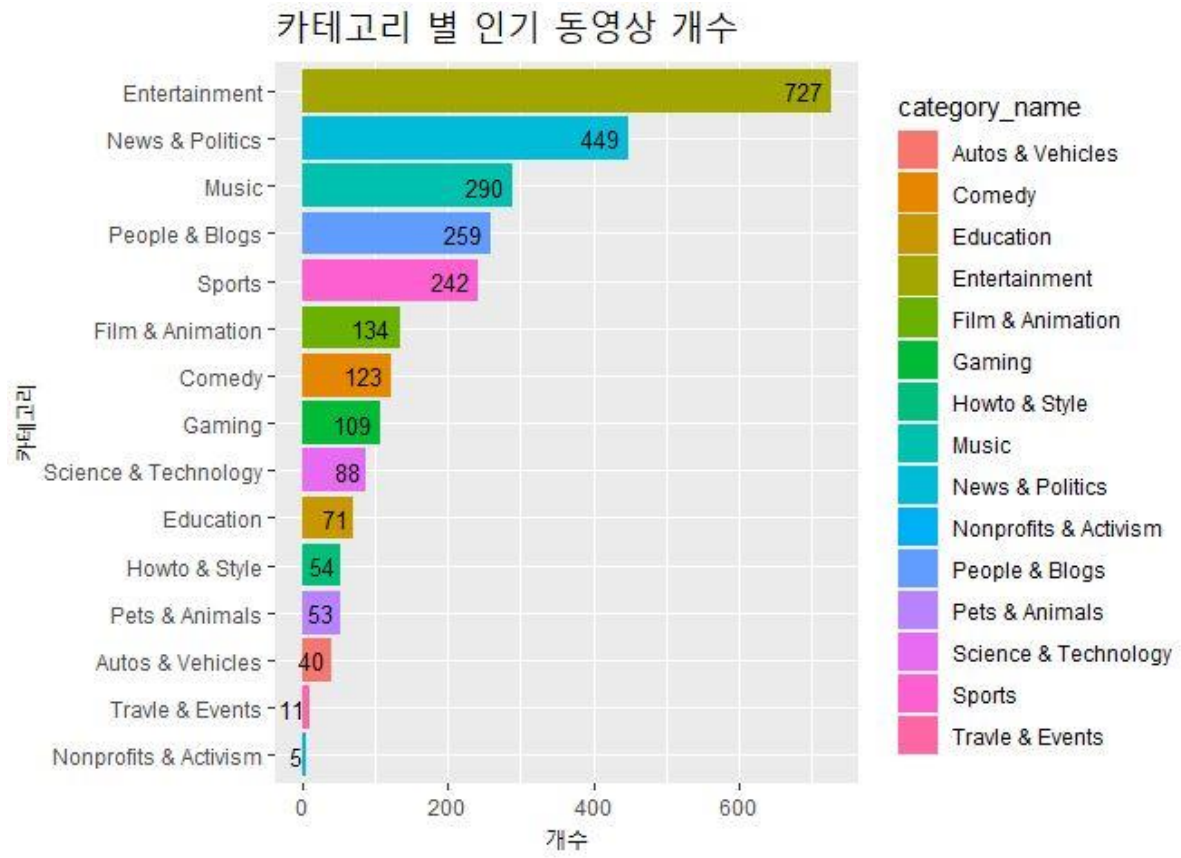
인기 동영상의 제목/태그/설명에 자주 사용되는 키워드는 무엇인가?

가설 4) 키워드 검색으로 인한 시청자 유입이 쉬워지므로 영상 설명 기재는 인기동영상 등재에 영향을 미칠 것 이다.

설명 워드클라우드

인기동영상에 가장 많이 등재되는 카테고리는 무엇인가?

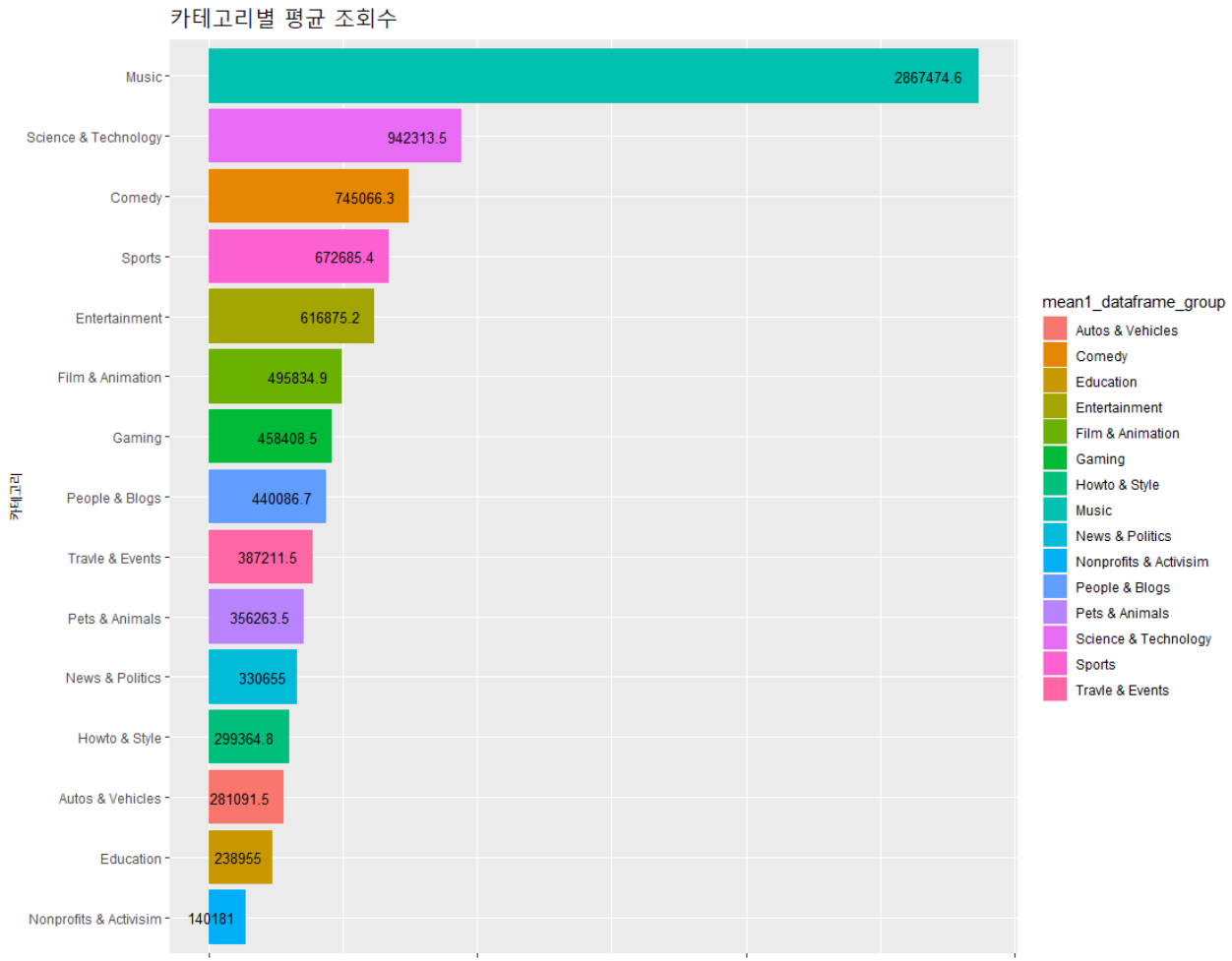
가설 5) 엔터테인먼트 혹은 게임으로 예상한다.



카테고리 별 인기 동영상 개수 히스토그램

평균적으로 조회수가 가장 높은 카테고리는 무엇인가?

가설 6) 엔터테인먼트 혹은 게임으로 예상한다.

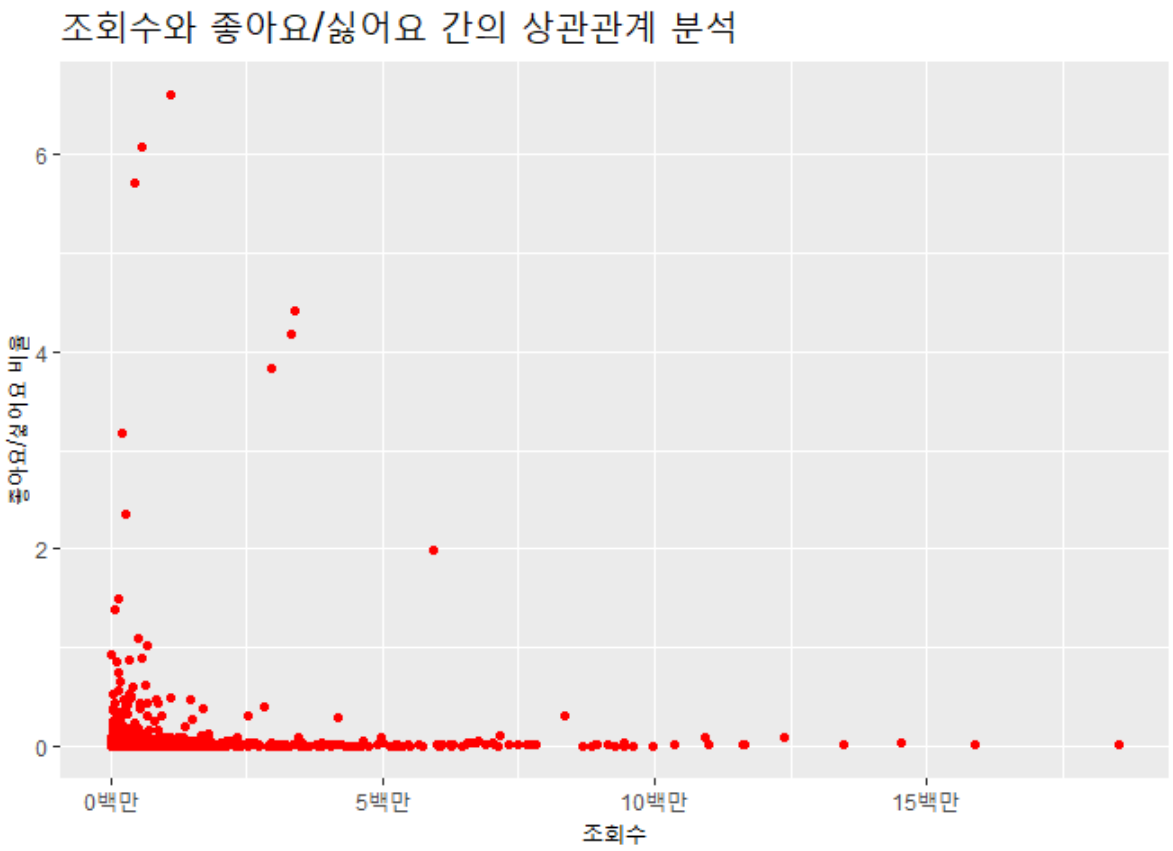


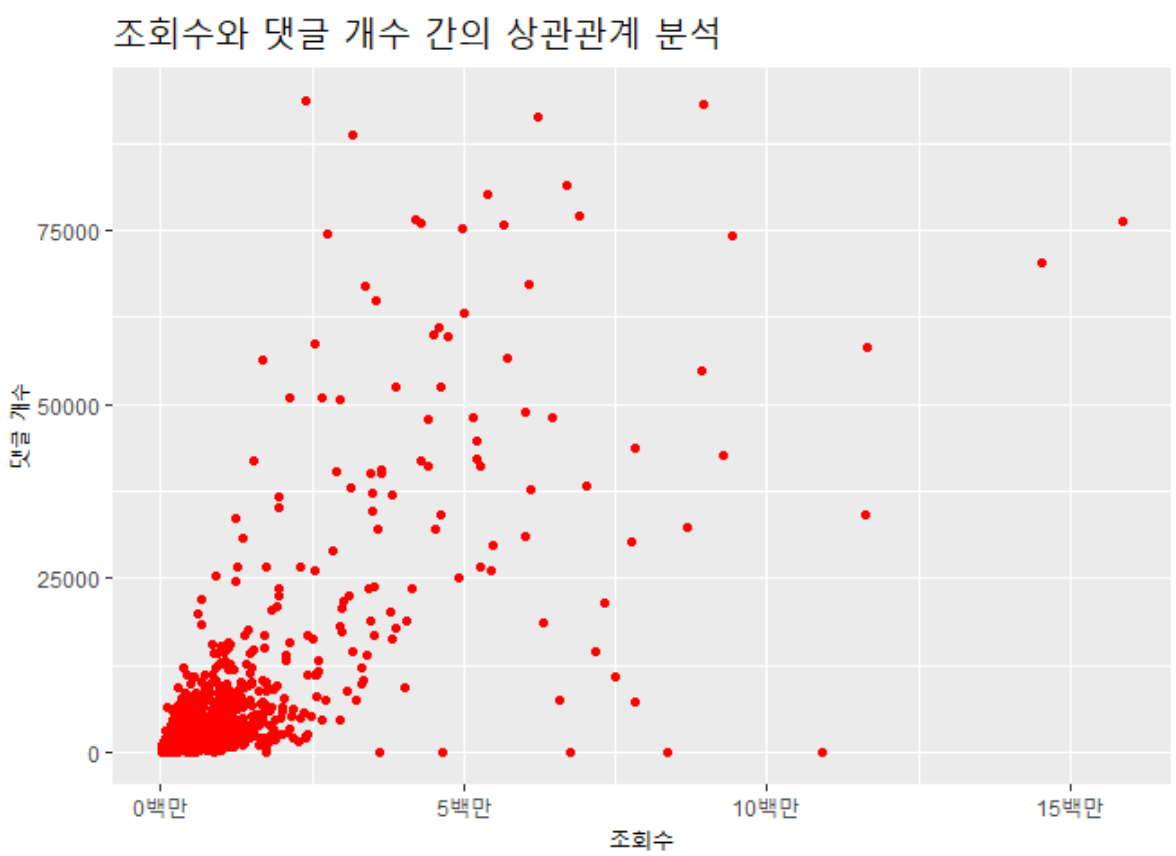
	category_name	mean_view_count
1	Music	2867474.6
2	Science & Technology	942313.5
3	Comedy	745066.3
4	Sports	672685.4
5	Entertainment	616875.2
6	Film & Animation	495834.9
7	Gaming	458408.5
8	People & Blogs	440086.7
9	Travle & Events	387211.5
10	Pets & Animals	356263.5
11	News & Politics	330655.0
12	Howto & Style	299364.8
13	Autos & Vehicles	281091.5
14	Education	238955.0
15	Nonprofits & Activism	140181.0

카테고리별 평균 조회수 히스토그램 & 데이터표

좋아요/싫어요, 댓글은 영상의 조회수에 영향을 미치는가?

가설 8) 인기 동영상 목록이나 키워드 검색 시 상단에 노출될 확률이 높아지기 때문에 영향을 미칠 것 이다.

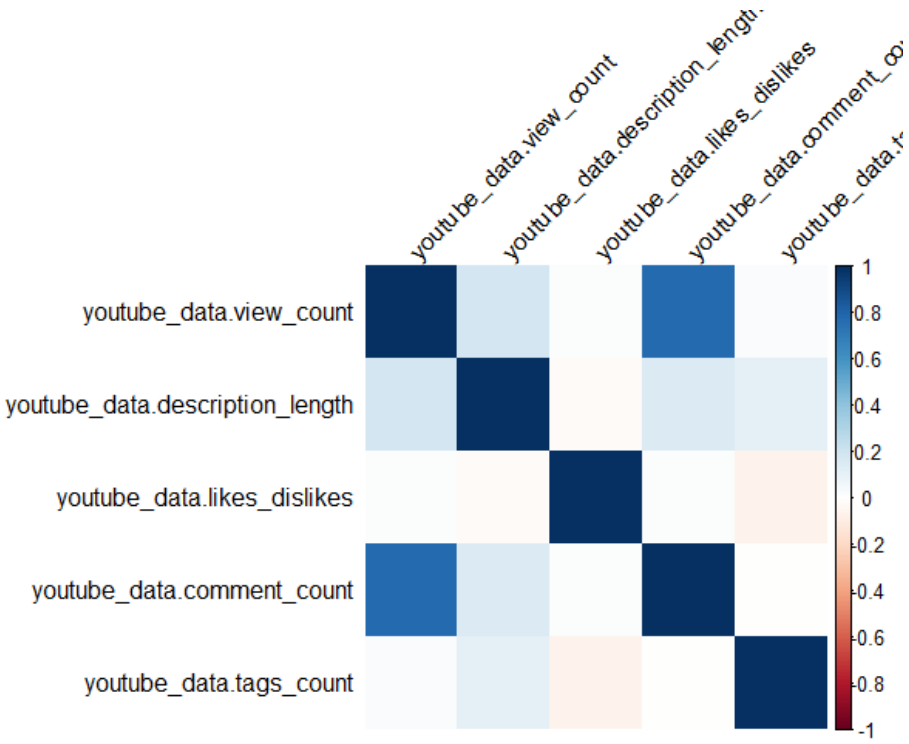




조회수와 좋아요/싫어요, 댓글 개수간 상관관계 산점도

좋아요/싫어요, 댓글은 영상의 조회수에 영향을 미치는가?

가설 8) 인기 동영상 목록이나 키워드 검색 시 상단에 노출될 확률이 높아지기 때문에 영향을 미칠 것 이다.



	youtube_data.view_count	youtube_data.description_length
youtube_data.view_count	1.00	0.19
youtube_data.description_length	0.19	1.00
youtube_data.likes_dislikes	0.01	-0.03
youtube_data.comment_count	0.77	0.15
youtube_data.tags_count	0.03	0.11

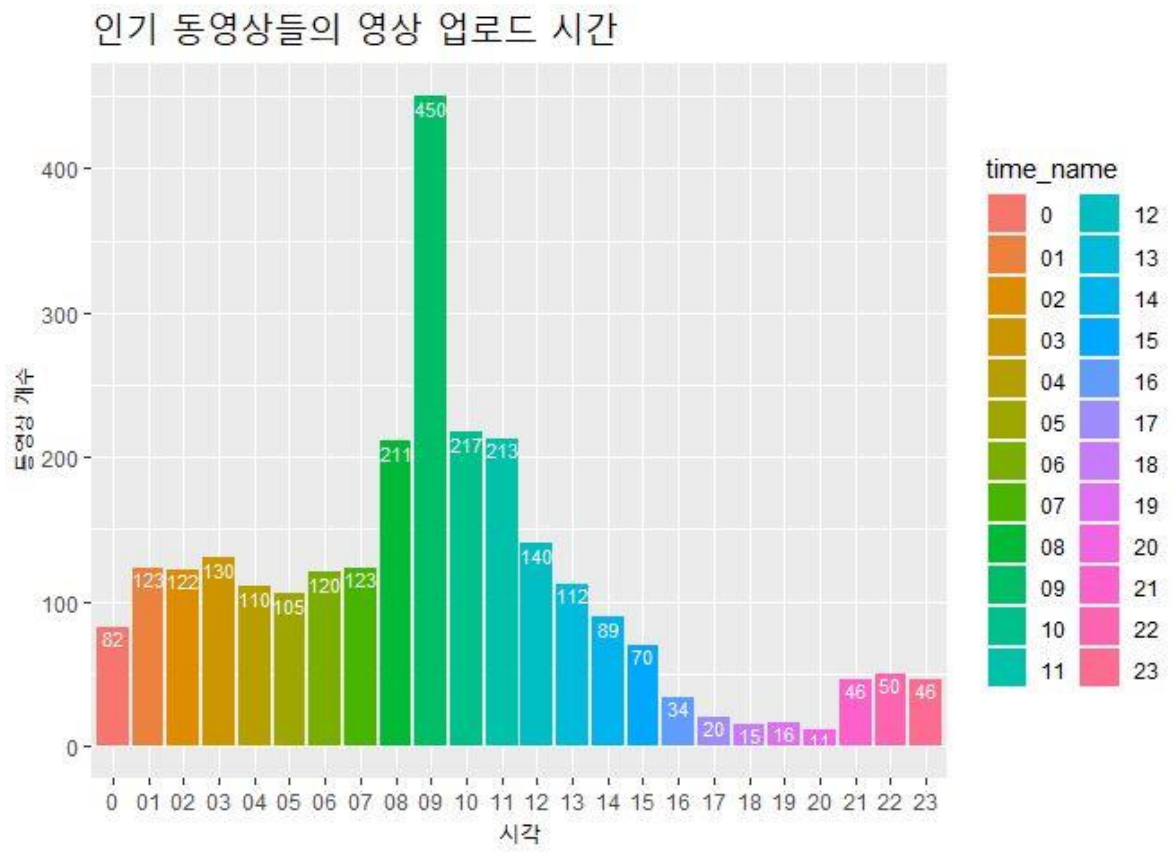
	youtube_data.likes_dislikes	youtube_data.comment_count
youtube_data.view_count	0.01	0.77
youtube_data.description_length	-0.03	0.15
youtube_data.likes_dislikes	1.00	0.01
youtube_data.comment_count	0.01	1.00
youtube_data.tags_count	-0.06	0.00

	youtube_data.tags_count
youtube_data.view_count	0.03
youtube_data.description_length	0.11
youtube_data.likes_dislikes	-0.06
youtube_data.comment_count	0.00
youtube_data.tags_count	1.00

각 항목들 상관계수 heatmap으로 출력

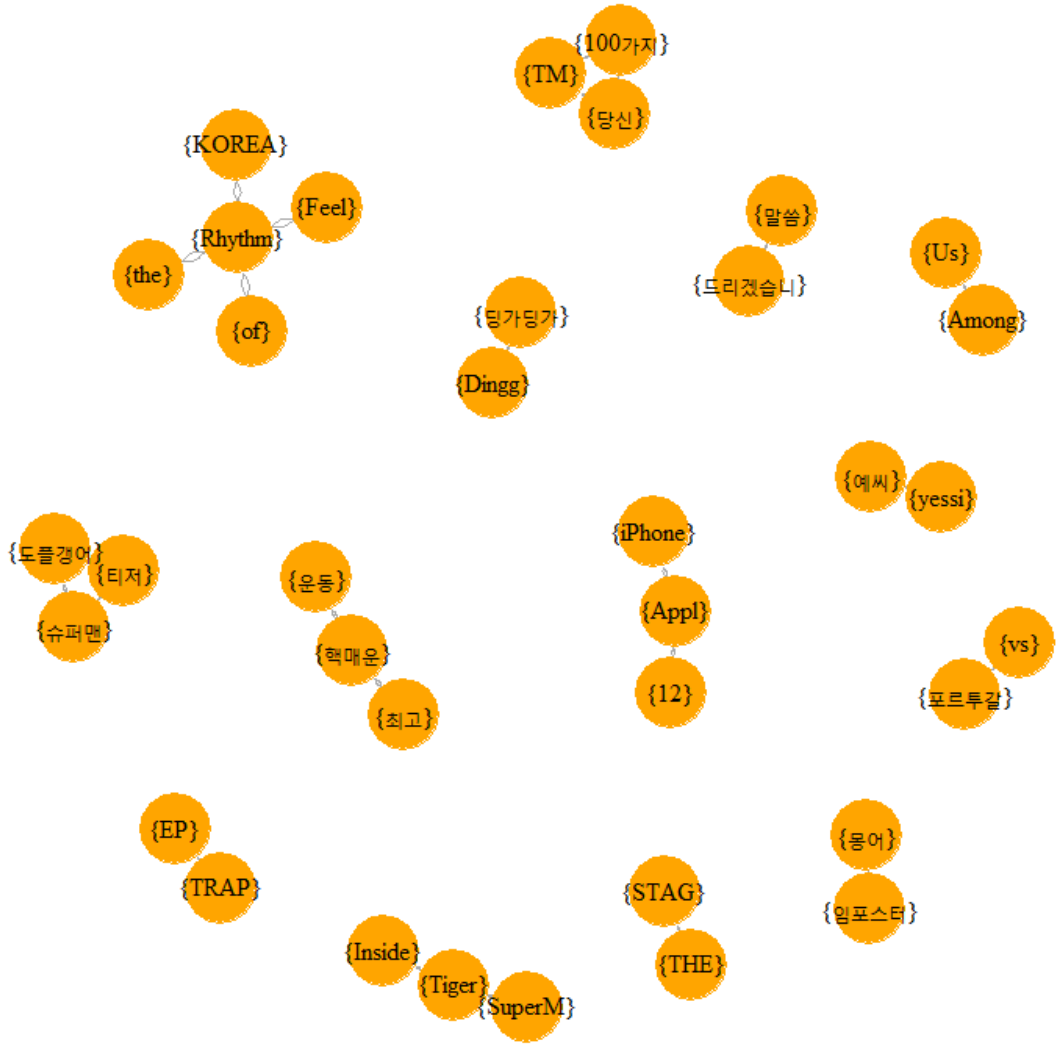
인기 동영상에 등재된 영상들의 업로드 시간은 어떠한가?

가설 9) 하교, 퇴근 후에 여가시간을 이용하여 영상을 시청하는 사람이 많으므로, 오후 6시 이후 업로드량이 가장 많을 것이다.



영상 업로드 시간 히스토그램

인기동영상 키워드 연관관계 분석



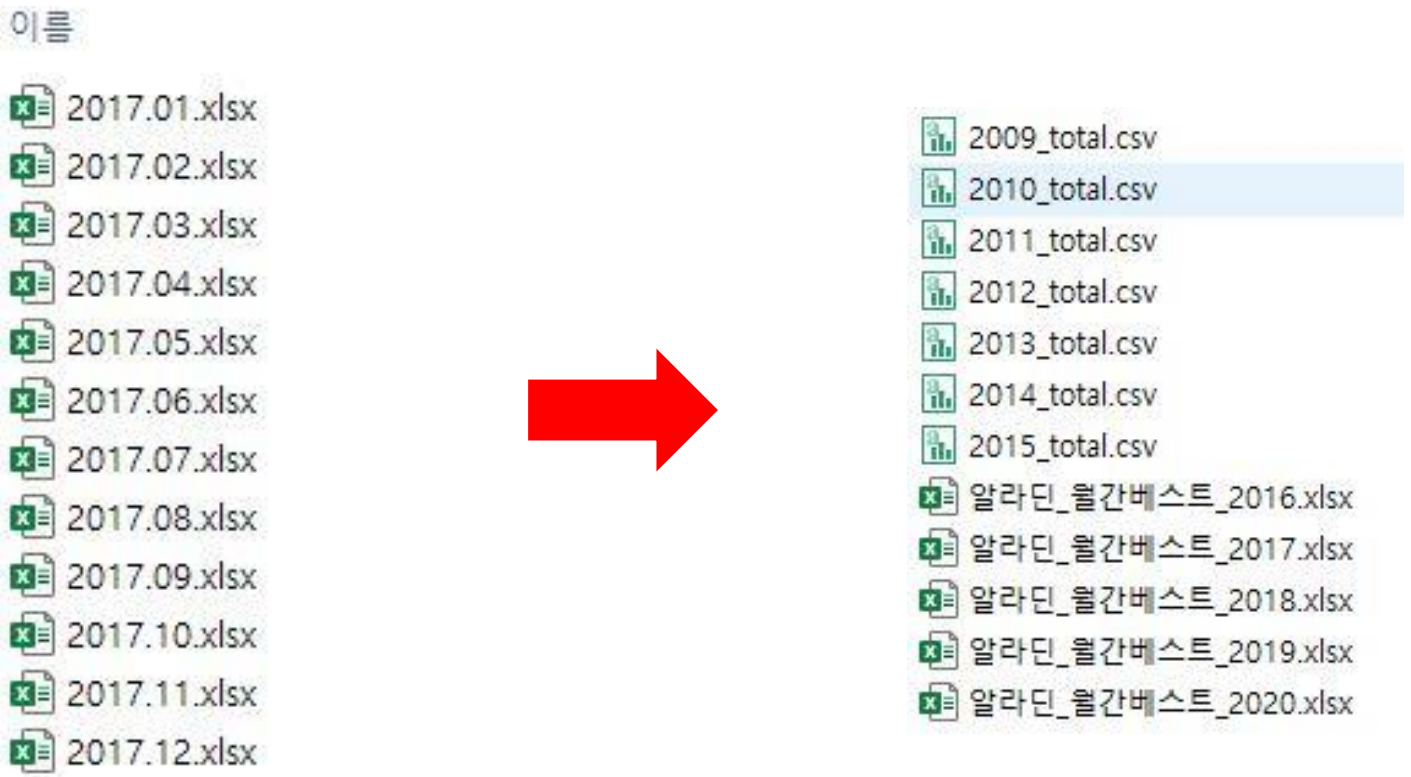
인기동영상 제목에 사용된 단어들 연관성 시각화

[12] {드리겠습니} => {말씀}
[13] {말씀} => {드리겠습니}
[14] {Dingg} => {딩가딩가}
[15] {딩가딩가} => {Dingg}
[16] {TRAP} => {EP}
[17] {EP} => {TRAP}
[18] {100가지} => {TM}
[19] {TM} => {100가지}
[20] {100가지} => {당신}
[21] {당신} => {100가지}
[22] {TM} => {당신}
[23] {당신} => {TM}
[24] {도플갱어} => {슈퍼맨}
[25] {슈퍼맨} => {도플갱어}
[26] {도플갱어} => {티저}
[27] {티저} => {도플갱어}
[28] {Appl} => {iPhone}
[29] {iPhone} => {Appl}
[30] {Appl} => {12}
[31] {12} => {Appl}
[32] {Us} => {Among}
[33] {Among} => {Us}
[34] {Tiger} => {Inside}
[35] {Inside} => {Tiger}
[36] {Tiger} => {SuperM}
[37] {SuperM} => {Tiger}
[38] {STAG} => {THE}
[39] {THE} => {STAG}
[40] {예씨} => {yessi}
[41] {yessi} => {예씨}
[42] {임포스터} => {몽어}
[43] {몽어} => {임포스터}
[44] {포르투갈} => {vs}
[45] {vs} => {포르투갈}
[46] {슈퍼맨} => {티저}
[47] {티저} => {슈퍼맨}
[48] {핵매운} => {운동}
[49] {운동} => {핵매운}
[50] {핵매운} => {최고}
[51] {최고} => {핵매운}
[52] {Rhythm} => {KOREA}
[53] {KOREA} => {Rhythm}
[54] {Rhythm} => {Feel}
[55] {Feel} => {Rhythm}
[56] {Rhythm} => {the}
[57] {the} => {Rhythm}
[58] {Rhythm} => {of}
[59] {of} => {Rhythm}
[60] {voyag} => {Bon}
[61] {Bon} => {voyag}
[101] {에이} => {트웬티} 0.0
[102] {트웬티} => {EP} 0.0
[103] {EP} => {트웬티} 0.0
[104] {ALIEN} => {LEE} 0.0
[105] {LEE} => {ALIEN} 0.0
[106] {ALIEN} => {SUHYUN}
[107] {SUHYUN} => {ALIEN}
[108] {Who} => {정명규} 0.
[109] {정명규} => {Who} 0.
[110] {Who} => {임영} 0.0
[111] {임영} => {Who} 0.0
[112] {revie} => {아이폰12} 0.0
[113] {아이폰12} => {revie} 0.0
[114] {UEL} => {토트넘} 0.0
[115] {토트넘} => {UEL} 0.0
[116] {UEL} => {21} 0.00
[117] {21} => {UEL} 0.00
[118] {UEL} => {20} 0.00
[119] {20} => {UEL} 0.00
[120] {UEL} => {vs} 0.00
[121] {vs} => {UEL} 0.00
[122] {공소시효} => {밍꼬발} 0.
[123] {밍꼬발} => {공소시효} 0.
[124] {밀리터리버거} => {롯데리아}
[125] {롯데리아} => {밀리터리버거}

인기동영상 제목에 사용된 단어들 연관성 시각화

알라딘 데이터 분석
시각화 작업 결과물

2. 출판 분야 데이터 전처리



월별 데이터를 연간 데이터로 합침

2. 출판 분야 데이터 전처리

순번	구분	상품명	ISBN	ISBN13	부가기호	출판사/제	저자/아티	정가	판매가	할인액	할인율	마일리지	출간일	세일즈포인트
1	그래픽/멀	맛있는 디	896848208	9.79E+12	13000	한빛미디어	윤이사라.부	22,000	19,800	2,200	10%	1,100점	20150805	1154
2	그래픽/멀	TCG 일러	896088166	9.79E+12	13000	아이생각(I	노진 지음	28,000	25,200	2,800	10%	1,400점	20150830	3286
3	그래픽/멀	스케치업 2	8.96E+09	9.79E+12	13000	정보문화사	한정훈 지	35,000	31,500	3,500	10%	1,750점	20151210	1472
4	그래픽/멀	만화로 배	8.93E+09	9.79E+12	13000	성안당	임완규 지	18,000	16,200	1,800	10%	900점	20151012	1088
5	그래픽/멀	포토샵 +	8.98E+09	9.79E+12	18000	혜지원	정윤선(윤	23,000	20,700	2,300	10%	1,150점	20140325	2260
6	그래픽/멀	맛있는 디	8.97E+09	9.79E+12	13000	한빛미디어	이수정 지	24,000	21,600	2,400	10%	1,200점	20141130	1922
7	그래픽/멀	파워포인트	8.93E+09	9.79E+12	13000	예문사	이혜강 지	18,000	16,200	1,800	10%	900점	20150510	1401
8	그래픽/멀	웹툰 스케	8.96E+09	9.79E+12	13650	아이생각(I	엘프화가	35,000	31,500	3,500	10%	1,750점	20160110	440
9	그래픽/멀	맛있는 디	8.97E+09	9.79E+12	13000	한빛미디어	윤성우.김	23,000	20,700	2,300	10%	1,150점	20150920	1450
10	그래픽/멀	배경 일러	8.97E+09	9.79E+12	13000	프리렉	garnet 지	20,000	18,000	2,000	10%	1,000점	20141115	893
11	그래픽/멀	The Game	K5224342	9.79E+12	1	비엘북스	김무광 외	35,000	31,500	3,500	10%	1,750점	20151228	337
12	그래픽/멀	편집 디자	8.97E+09	9.79E+12	13000	한빛미디어	황지완 지	25,000	22,500	2,500	10%	1,250점	20140420	1599
13	그래픽/멀	포토샵 디	8.97E+09	9.79E+12	13000	한빛미디어	정다영 지	25,000	22,500	2,500	10%	1,250점	20150130	594
14	그래픽/멀	인디자인 (	9E+09	9.79E+12	1	채움북스	윤고선 지	35,000	31,500	3,500	10%	1,750점	20141010	1110
15	그래픽/멀	꿈돌이의	8.99E+09	9.79E+12	13000	청담북스	권경범 지	30,000	27,000	3,000	10%	1,500점	20140830	1289
16	그래픽/멀	꿈돌이의	8.99E+09	9.79E+12	13000	청담북스	권경범 지	32,000	28,800	3,200	10%	1,600점	20150120	858
17	그래픽/멀	일러스트	9E+09	9.79E+12	13600	CABOOKS	CA 편집부	22,000	20,900	1,100	5%	660점	20140526	743
18	그래픽/멀	웰스트리트	8.97E+09	9.79E+12	13320	인사이트	도나 M. 웰	15,000	13,500	1,500	10%	750점	20140314	3031
19	그래픽/멀	포토샵 CS	896030333	9.79E+12	13000	황금부엉이	송병용.주	22,800	20,520	2,280	10%	1,140점	20130123	1557
20	그래픽/멀	AutoCAD	K4124341(	9.79E+12	3000	길벗	권현실 지	30,000	27,000	3,000	10%	1,500점	20160104	636
21	그래픽/멀	손맵 핸드	K6324336	9.79E+12	13000	비엘북스	윤상혁 지	36,000	32,400	3,600	10%	1,800점	20151015	709
22	그래픽/멀	포토샵 아	9E+09	9.79E+12	13600	CABOOKS	CA 편집부	22,000	20,900	1,100	5%	660점	20150707	624
23	그래픽/멀	정영헌의	8.96E+09	9.79E+12	93560	아이생각(I	정영헌 지	45,000	40,500	4,500	10%	2,250점	20150905	337
24	그래픽/멀	스콧 켈비	896848032	9.79E+12	13000	한빛미디어	스콧 켈비	23,000	20,700	2,300	10%	1,150점	20130712	981
25	그래픽/멀	비주얼 이	K9424330	9.79E+12	13	비엘북스	Ron Ganb	36,000	32,400	3,600	10%	1,800점	20150827	763

Before

2. 출판 분야 데이터 전처리

순번	상품명	출간일	best_date
1	맛있는 디자인 포토샵 CC	20150805	201601
2	TCG 일러스트 작법서 입문편	20150830	201601
3	스케치업 2015 & V-Ray	20151210	201601
4	만화로 배우는 클립스튜디오	20151012	201601
5	포토샵 + 일러스트레이터 작업의 기술	20140325	201601
6	맛있는 디자인 애프터이펙트 CS6 & CC	20141130	201601
7	파워포인트 for 인포그래픽	20150510	201601
8	웹툰 스케치업	20160110	201601
9	맛있는 디자인 프리미어 프로 CS6 & CC	20150920	201601
10	배경 일러스트 테크닉	20141115	201601
11	The Game Graphics : 유니티와 언리얼 그리고 VR	20151228	201601
12	편집 디자인 강의 + 인디자인	20140420	201601
13	포토샵 디자인 강의	20150130	201601
14	인디자인 CC/CS6 슈퍼테크닉 노트	20141010	201601
15	꿈돌이의 라이노 5 : Rhino 3D 프린터의 첫걸음	20140830	201601
16	꿈돌이의 라이노 5 : Rhino 3D 곡면 모델링의 원리와 기법	20150120	201601
17	일러스트레이터 아트웍	20140526	201601
18	월스트리트저널 인포그래픽 가이드	20140314	201601
19	포토샵 CS6 Using Bible	20130123	201601
20	AutoCAD 2016 무작정 따라하기	20160104	201601

After

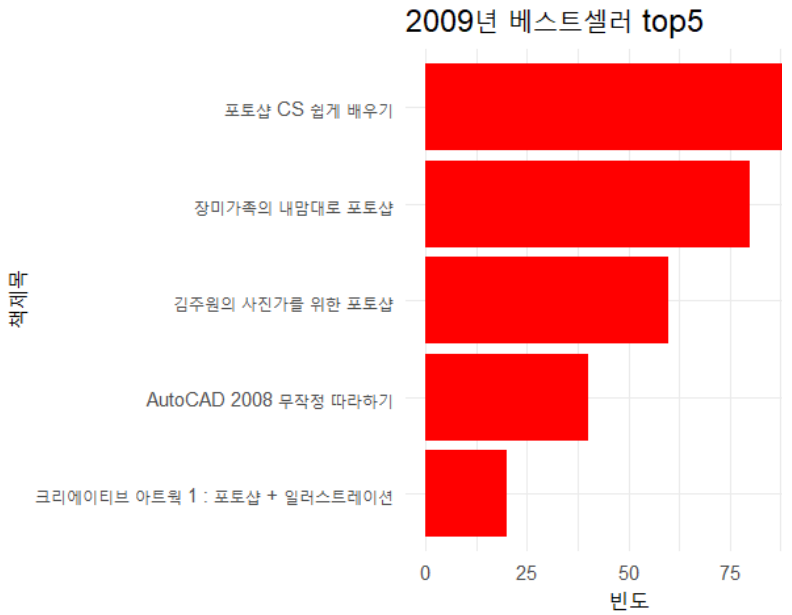
2. 출판 분야 워드 클라우드

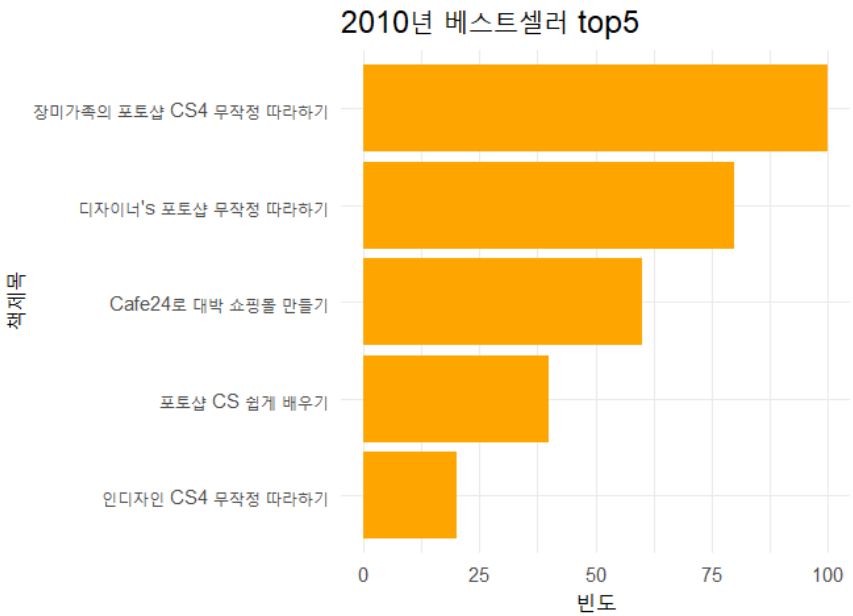
It/그래픽/멀티미디어 분야

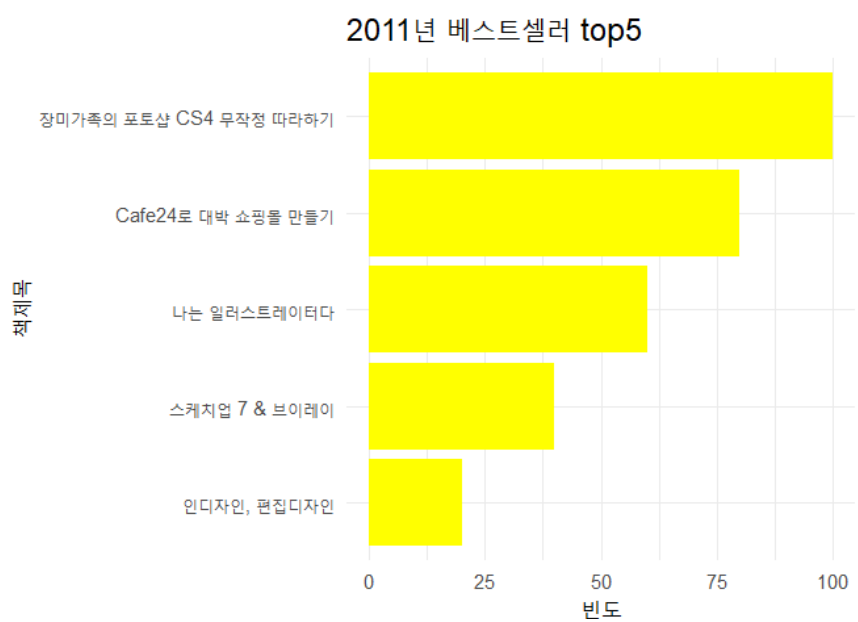
2010년

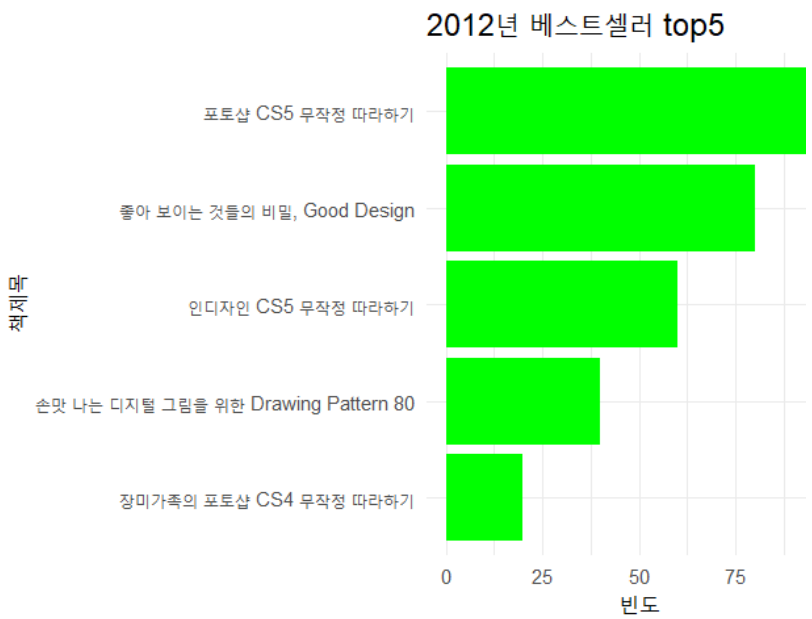
2020년

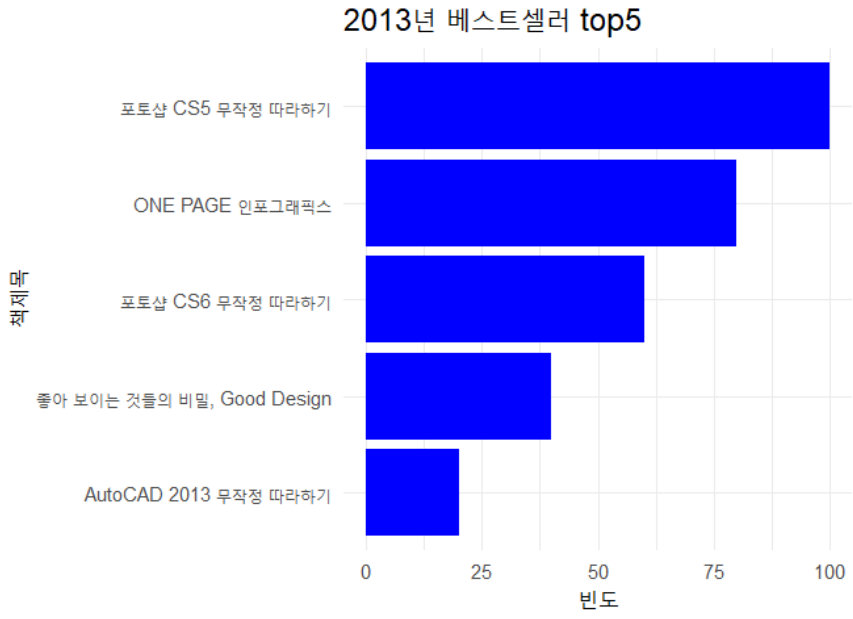
2. 출판 분야 막대 그래프 - 2009년~2014년 (it/그래픽/멀티미디어 분야)

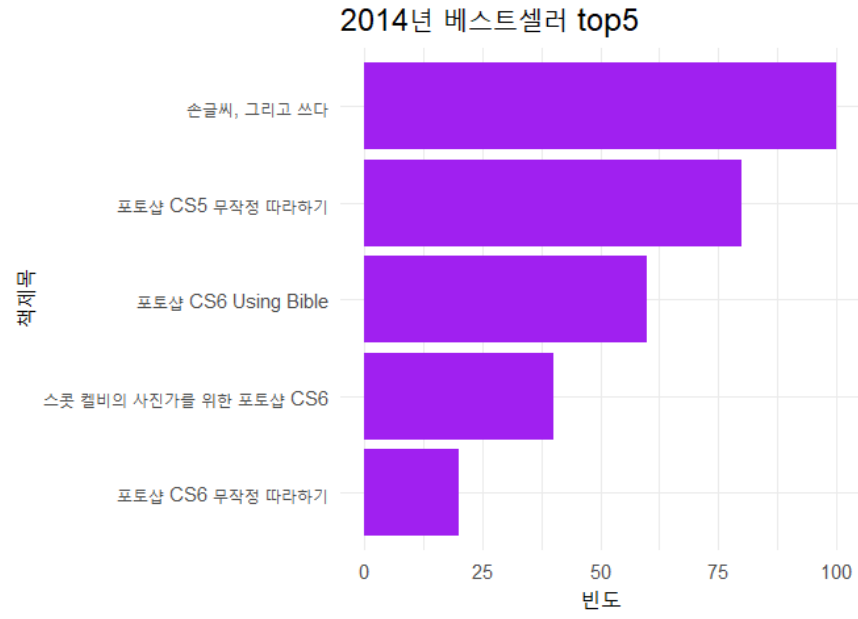




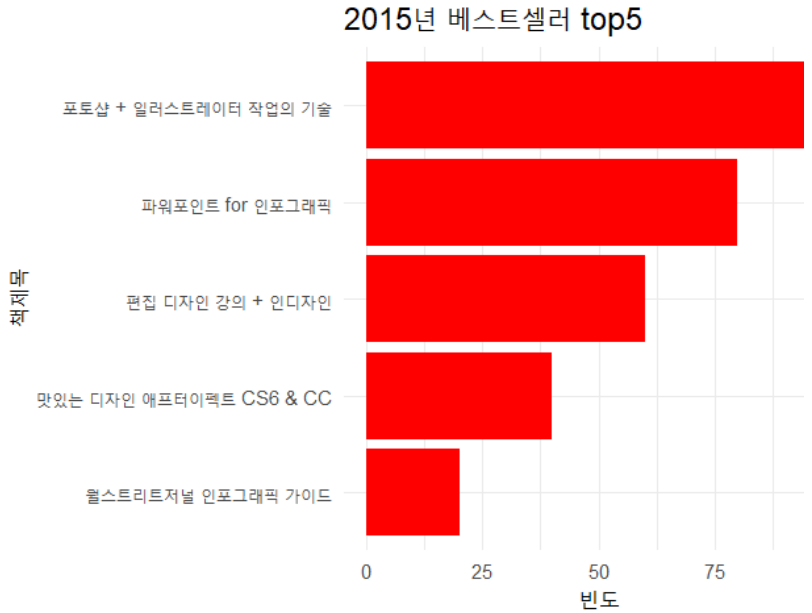


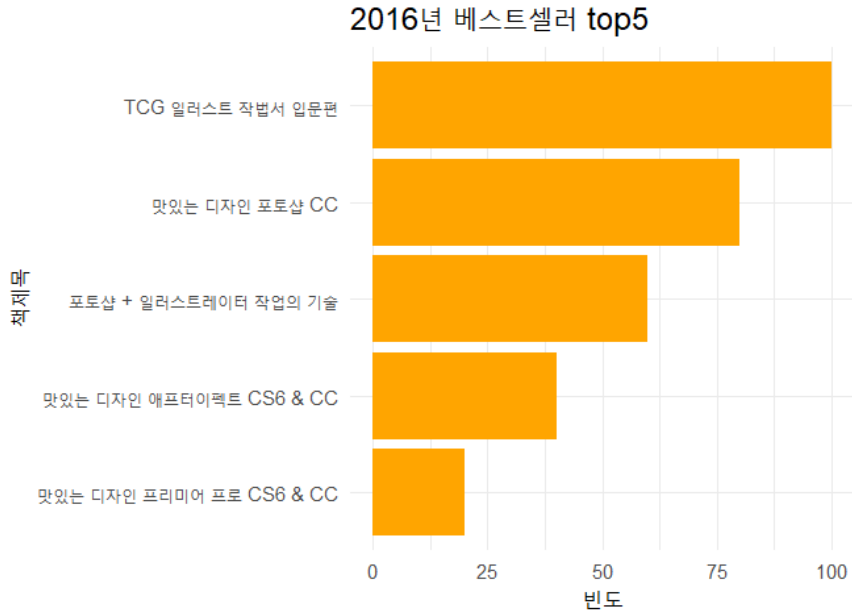


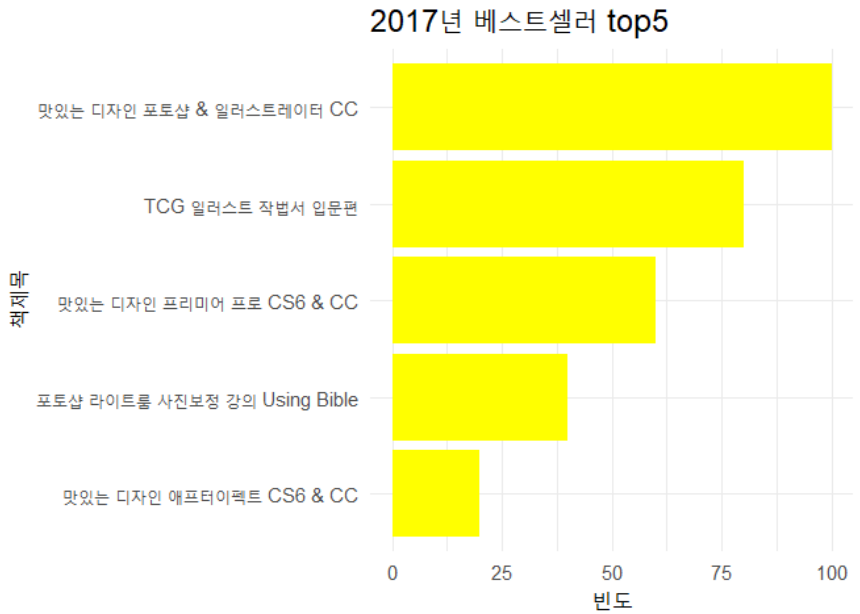


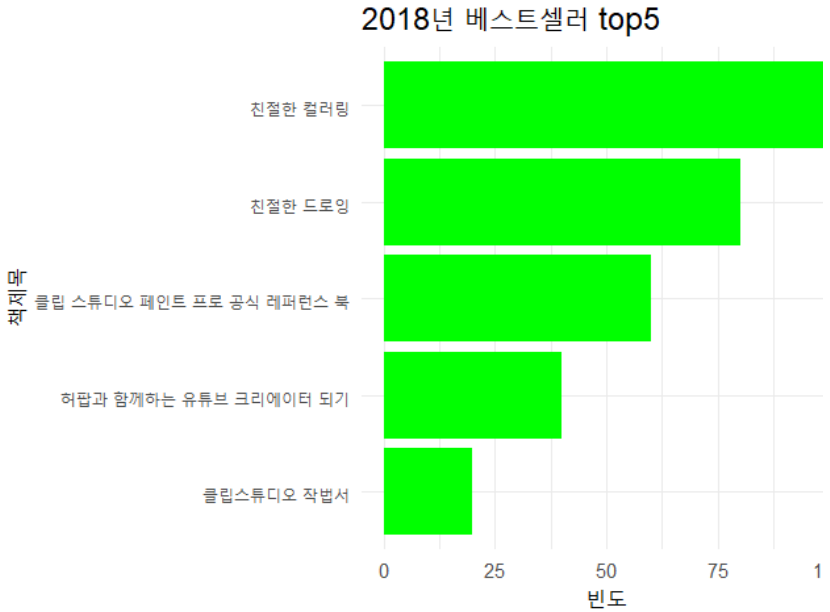


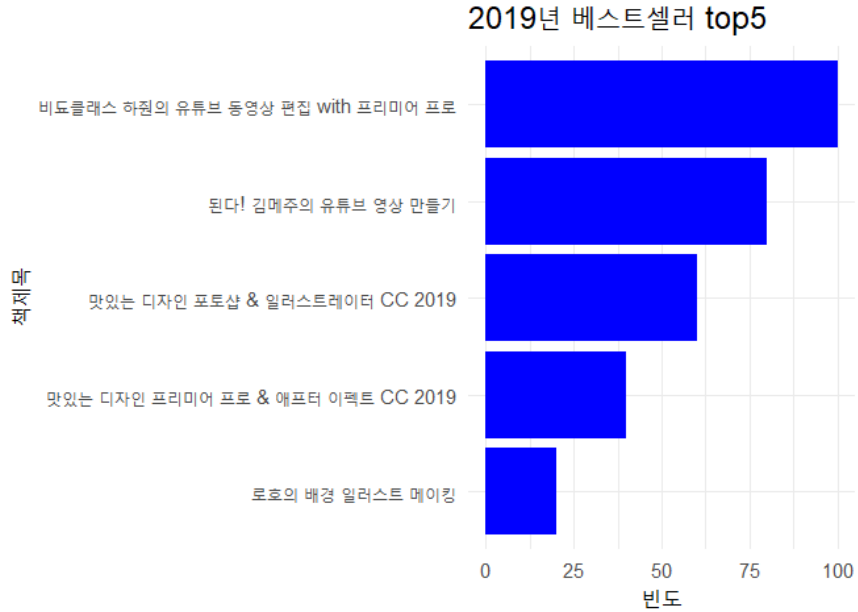
2. 출판 분야 막대 그래프 - 2015년 ~ 2020년 11월 (it/그래픽/멀티미디어 분야)

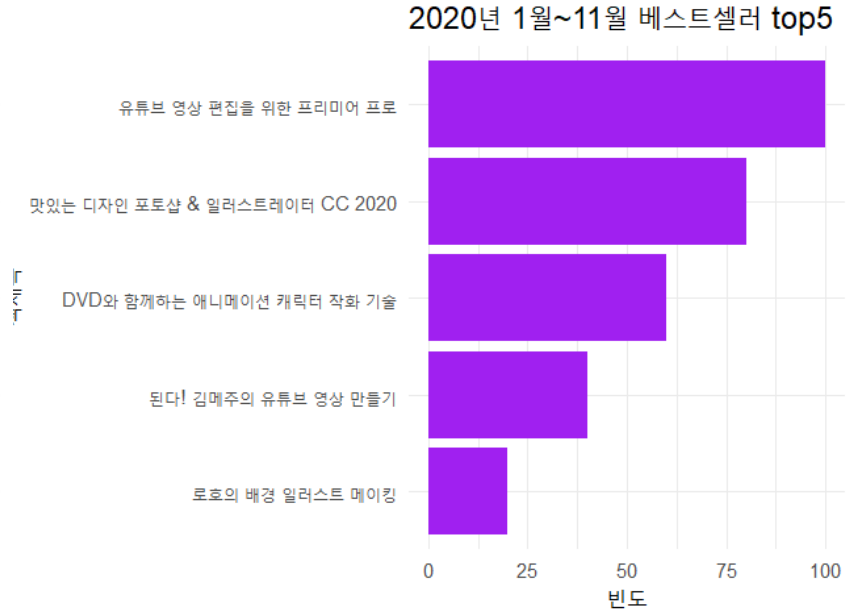




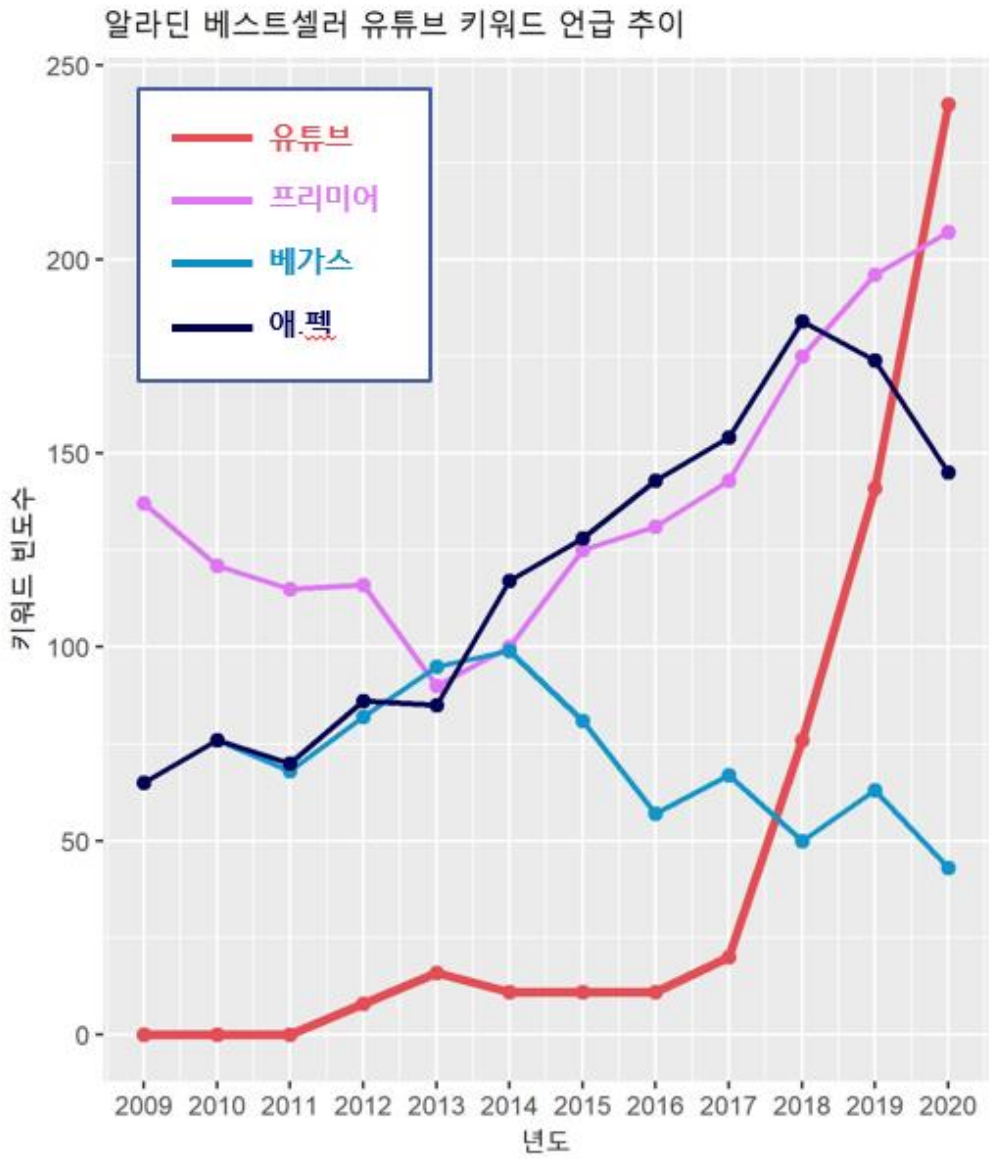








2. 출판 분야 꺾은선 그래프



유튜브 키워드 언급 횟수

2017년을 기점으로 가파른 상승세