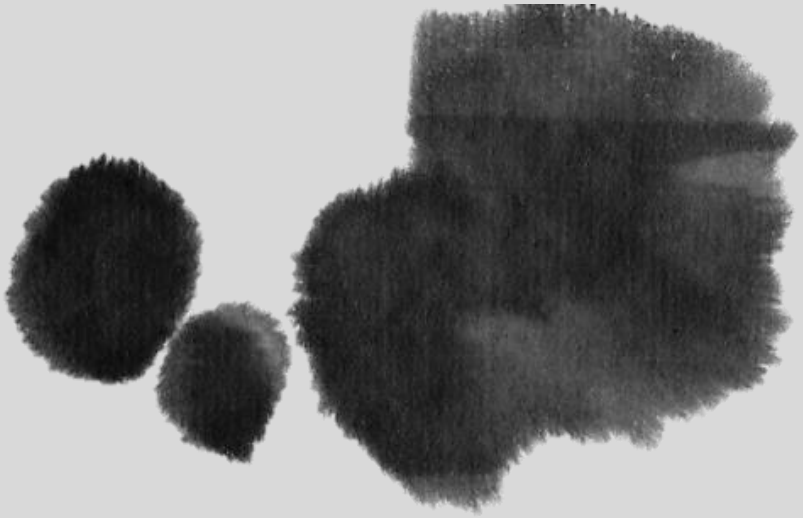
웹 크롤링을 활용한 유직 차트 시각화

백광흠

목 차



1. 웹 크롤링이란?

웹크롤링 이란?

“ 웹 크롤링이란 컴퓨터 소프트웨어 기술로

웹 사이트들에서 원하는 정보를 추출하는 것 “

크롤러(crawler)란 기어가는 사람 혹은 포복동물 이라는 의미로, 조직적, 자동적인 방법으로 각종 웹 페이지들을 돌아다니며 웹 문서의 URL, 링크정보, 문서내용 등 다량의 정보들을 수집해 오는 기능으로 인해 이런 이름이 붙었습니다.

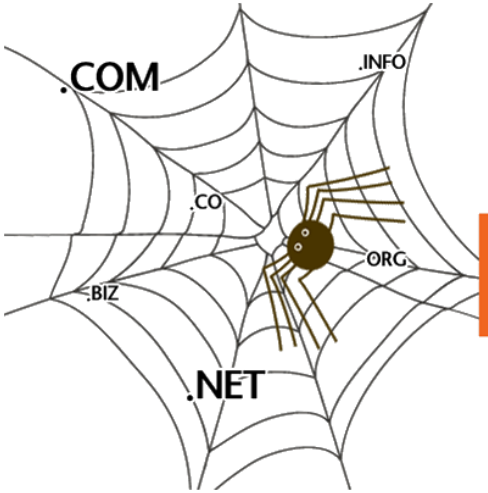
웹 크롤러(Web Crawler)는 방대한 웹 페이지를 두루두루 방문하여, 각종 정보를 자동적으로 수집하는 일을 하는 프로그램으로서 검색 엔진의 근간이 됩니다.

웹크롤링 이란?

웹 크롤링

웹 크롤러가 하는 작업을 **웹 크롤링(web crawling)** 혹은 **스파이더링(spidering)** 이라고 부르기도 하는데, 검색 엔진과 같은 여러 사이트에서는 데이터의 최신 상태 유지를 위해 항상 웹 크롤링을 합니다.

웹 크롤러는 대체로 방문한 사이트의 모든 페이지의 복사본을 생성하는 데 사용됩니다.

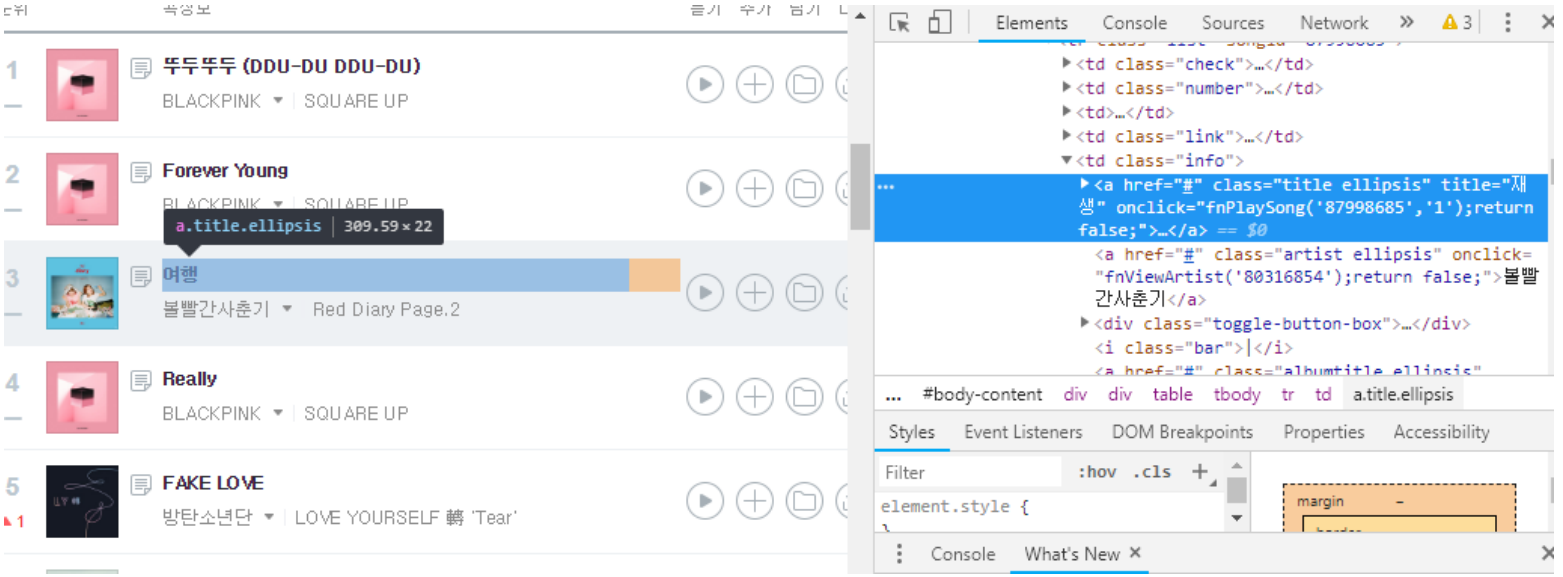


웹 크롤러Web Crawler
웹페이지를 방문하며
자동적으로 수집하는 프로그램

웹크롤링 이란?

웹은 기본적으로 HTML 형태로 되어 있습니다. 눈으로 볼 수 있다면, 해당 정보가 HTML 형태로 어떻게 보여지는지도 **'페이지 소스 보기'** 또는 **'개발자 검사'** 로 볼 수 있습니다.

이런 소스들은 개발자들이 정형화된 형태로 관리하고 있기 때문에 규칙이 생깁니다. 이런 규칙을 분석해서 우리가 **원하는 정보들만 뽑아오는 것**을 웹 크롤링 작업이라고 생각하면 됩니다.



웹크롤링 이란?

웹 크롤링

■ 웹 크롤러가 없다면?

예제 1) 음원 차트를 분석 및 가공 하기 위해 최근 3년 간의 음원 차트 데이터가 필요하다!

1. 멜론 홈페이지 접속
2. 2018.6.18 차트의 1~100위의 순위, 노래, 아티스트 등의 정보를 엑셀에 입력
3. 2018.6.17 차트의 1~100위의 순위, 노래, 아티스트 등의 정보를 엑셀에 입력

   ….
4. 2015.6.18 차트의 1~100위의 순위, 노래, 아티스트 등의 정보를 엑셀에 입력
5. 데이터 정제 시작

= 같은 작업을 수 백번 반복해야 한다.

2. 웹 크롤링 방법

웹 크롤링 방법

■ 웹 크롤링 절차

1. 패키지 설치
```
install.packages("XML")
install.packages("RCurl")
```

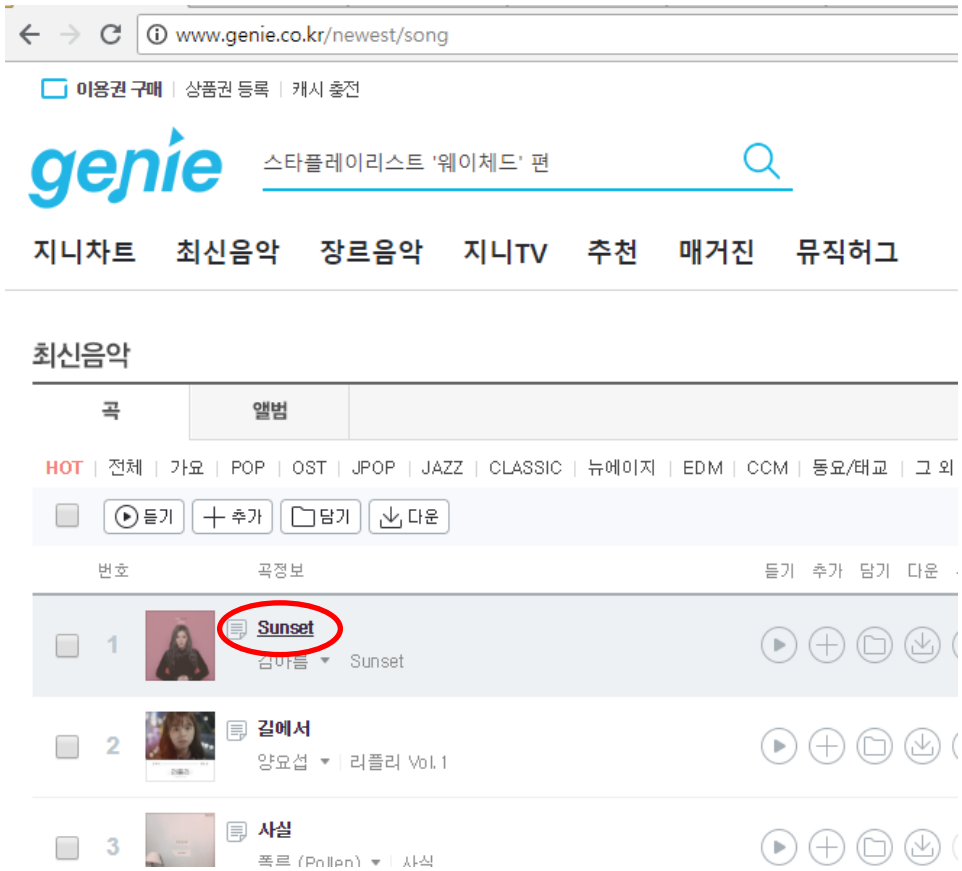
2. 필요한 데이터가 있는 url 주소를 가져온다.
```
url = "http://www.genie.co.kr/newest/song"
SOURCE = getURL(url)
PARSED = htmlParse(SOURCE)
```

3. Xpath 값과 xpathSApply 함수를 이용해 데이터를 가져온다.

* Xpath : XML 문서의 특정 요소나 속성에 접근하기 위한 경로를 지정하는 언어

웹 크롤링 방법

크롤링사용법



지니 홈페이지의 최신음악에서 첫 번째 노래 제목을 R로 가져오자.

웹 크롤링 방법

1. R에서 필요한 패키지를 설치 및 로드 한다.

```
install.packages("XML")
install.packages("RCurl")

library(XML)
library(RCurl)
```

2. 필요한 데이터가 있는 URL 주소를 가져온다.

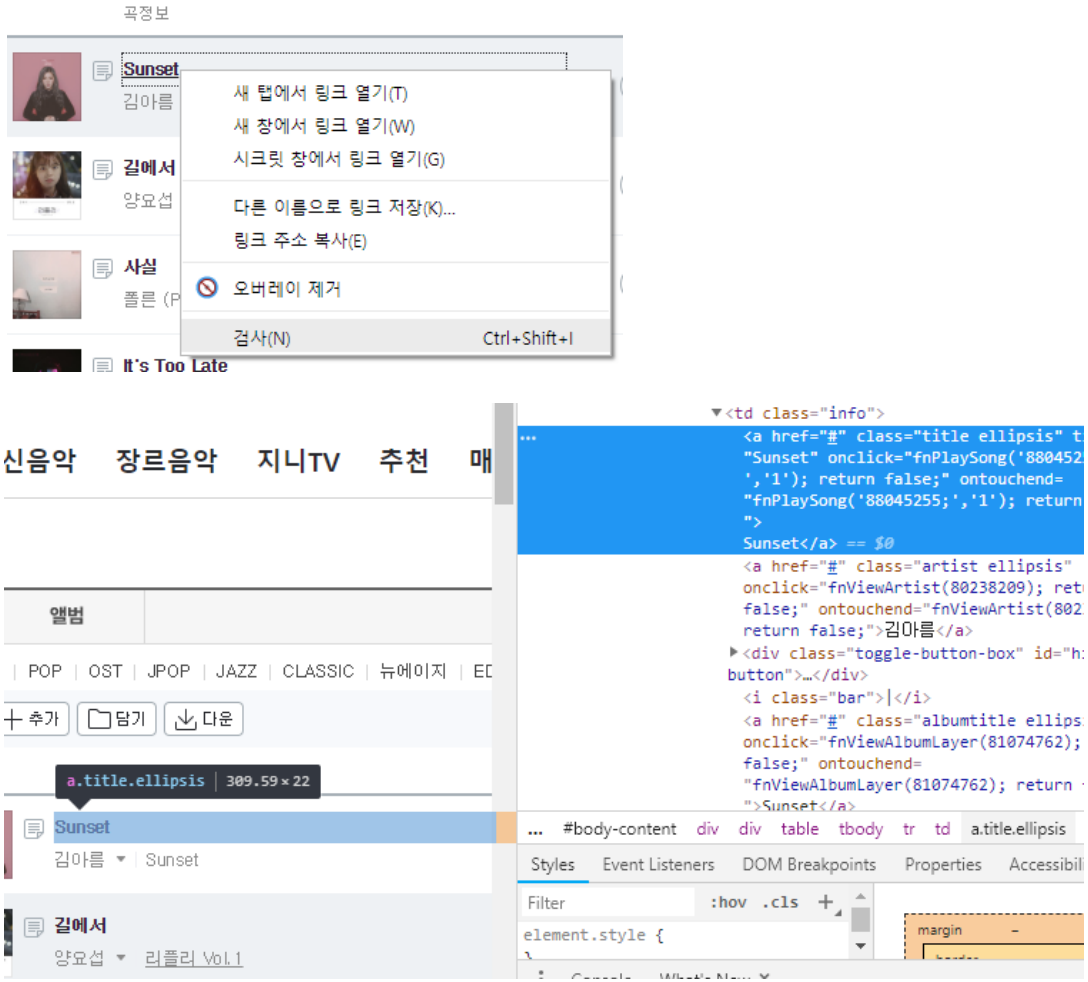
```
url = "http://www.genie.co.kr/newest/song"
SOURCE = getURL(url)
PARSED = htmlParse(SOURCE)
```

```
            $.ajax({
                type: "POST",
                url: "/myMusic/saveMyalbumList",
                dataType: "json",
                data: {  "unm":iMemUno , "maIds": maIds  },
                success: function (responseData) {
                    var retCode = responseData.Result.RetCode;
                    var retMSG  = responseData.Result.RetMsg;
                    if (retCode == 'A00001') {
                        alert ('챙혺?챙헹짜 챘헰헹챙헸헦챙헩뗍챘헹헦챘헹짚');
                        location.reload();
                    } else{
                        alert('챙혺?챙헹짜챙헸헰 챙헹짚챠헴짢챠헷헦챙헩뗍챘헹헦챘헹짚- '+ret
                    }
                }

            });

        }
    }

    //]]>
</script><!-- 챘짰짚챙짠혖챘쨔혈챘헴헴챙헹짚 챠헴헾챙헸혎 --><div class="layer-popup" styl
                                <div class="inner inspection">
                                    <h4>챘짰짚챙짠혖챘쨔혈챘헴헴챙헹짚</h4>
```

웹 크롤링 방법

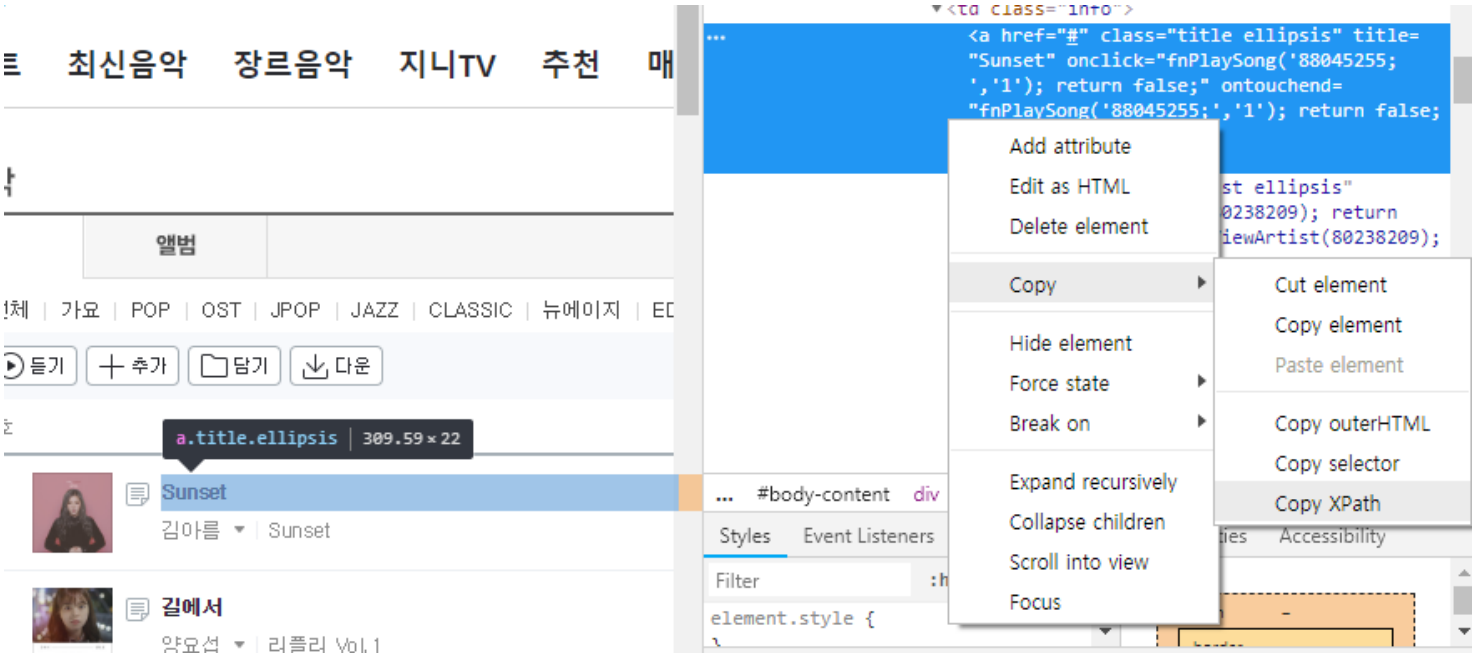
크롤링사용법

3. Xpath 값과 xpathSApply 함수를 이용해 데이터를 가져온다.



웹 크롤링 방법

크롤링사용법



//*[@id="body-content"]/div[4]/div/table/tbody/tr[1]/td[5]/a[1] 란 값이 복사된다.

웹 크롤링 방법


```
xpathSApply(PARSED,"//*[@id='body-content']/div[4]/div/table/tbody/tr[1]/td[5]/a[1]",xmlValue)
```

```
> xpathSApply(PARSED,"//*[@id='body-content']/div[4]/div/table/tbody/tr[1]/td[5]/a[1]",xmlValue)
[1] "\r\nSunset"
```

```
▶ <td class="link">…</td>
▼ <td class="info">
    <a href="#" class="title ellipsis" title="Sunset" onclick="fnPlaySong('88045255;','1');
    return false;" ontouchend="fnPlaySong('88045255;','1'); return false;">
    Sunset</a> == $0
    <a href="#" class="artist ellipsis" onclick="fnViewArtist(80238209); return false;"
    ontouchend="fnViewArtist(80238209); return false;">김아름</a>
```

■ 속성값 가져오기

```
    <a href="#" class="title ellipsis" title="Sunset" onclick="fnPlaySong('88045255;','1');
    return false;" ontouchend="fnPlaySong('88045255;','1'); return false;">
    Sunset</a> == $0
```

```
node_title = function(node) xmlAttrs(node)["title"]   # title 속성
xpathSApply(PARSED,"//*[@id='body-content']/div[4]/div/table/tbody/tr[1]/td[5]/a[1]",node_title)
```

```
> node_title = function(node) xmlAttrs(node)["title"]   # title 속성
> xpathSApply(PARSED,"//*[@id='body-content']/div[4]/div/table/tbody/tr[1]/td[5]/a[1]",node_title)
   title
"Sunset"
```

웹 크롤링 방법

크롤링 사용법

■ 크롤링 예제 소스

```
install.packages("XML")
install.packages("RCurl")

library(XML)
library(RCurl)

url = "http://www.genie.co.kr/newest/song"
SOURCE = getURL(url)
PARSED = htmlParse(SOURCE)

xpathSApply(PARSED,"//*[@id='body-
content']/div[4]/div/table/tbody/tr[1]/td[5]/a[1]",xmlValue)

node_title = function(node) xmlAttrs(node)["title"]   # title 속성
xpathSApply(PARSED,"//*[@id='body-
content']/div[4]/div/table/tbody/tr[1]/td[5]/a[1]",node_title)
```

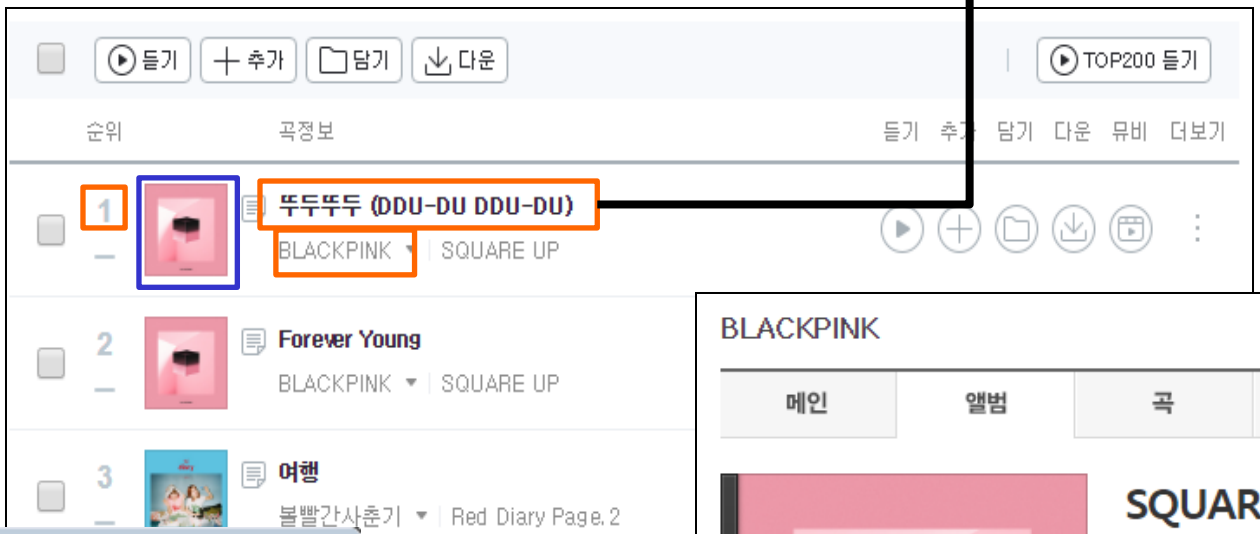
3. 뮤직차트 시각화

음원 차트 시각화

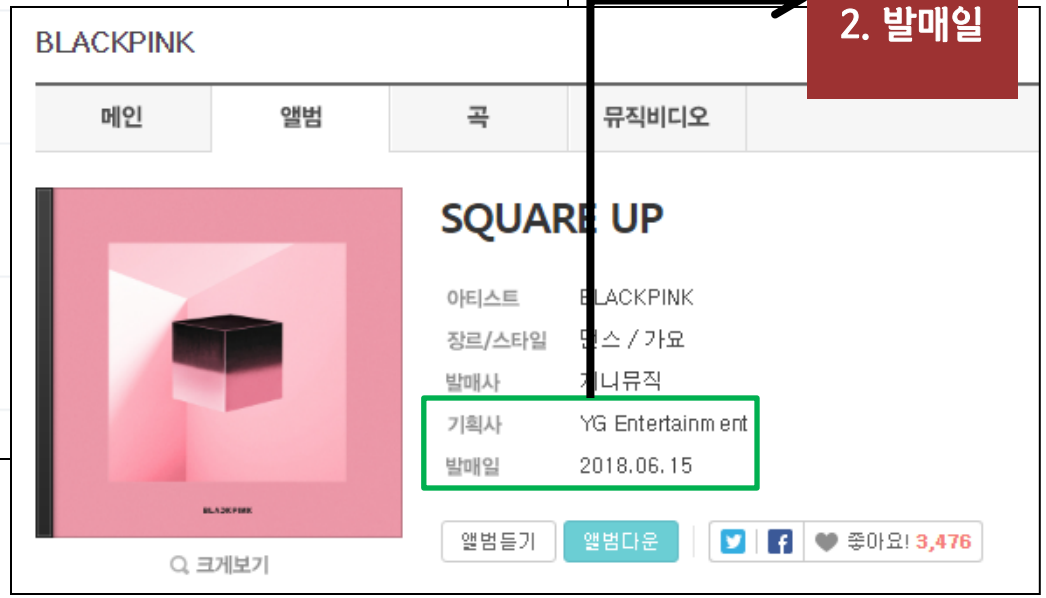
시각화

■ 크롤링 원리

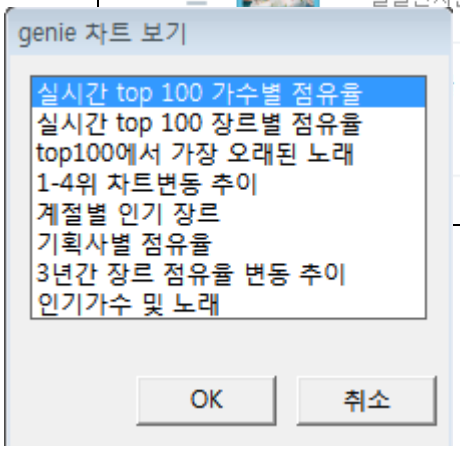
URL : https://www.genie.co.kr/chart/top200



1. 순위
2. 제목
3. 가수



1. 기획사
2. 발매일



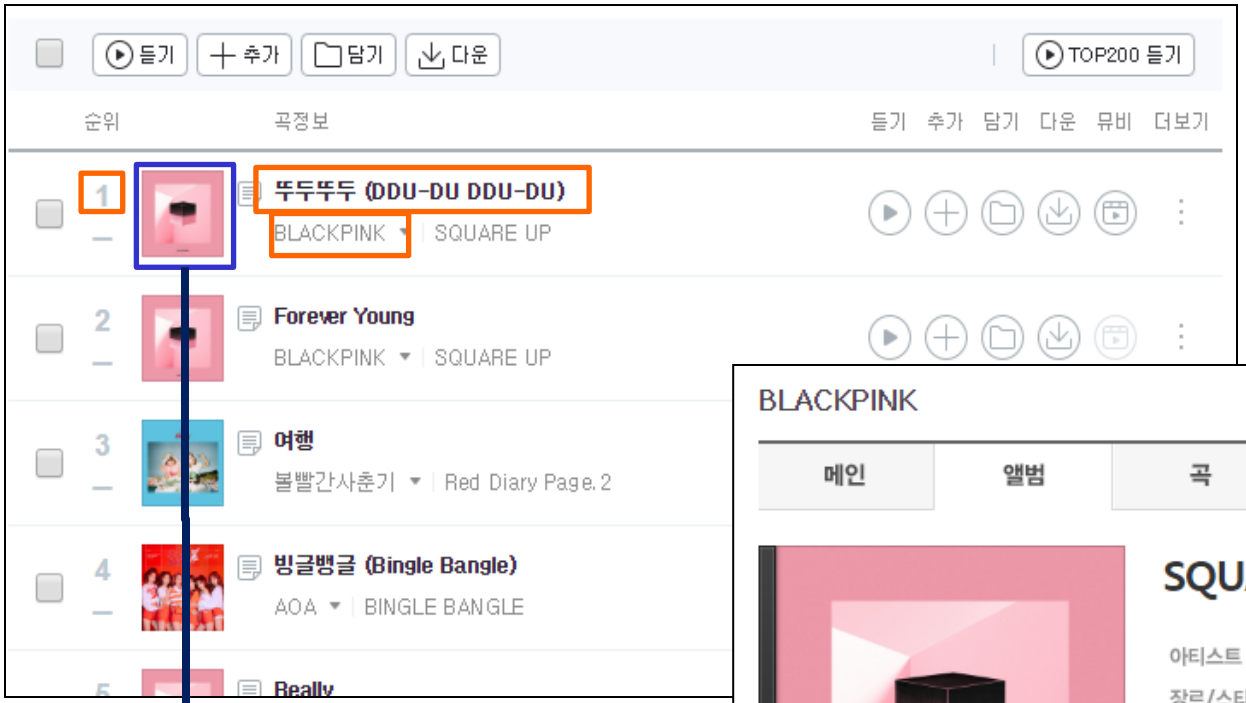
URL : https://www.genie.co.kr/detail/albumInfo?axnm=81074005

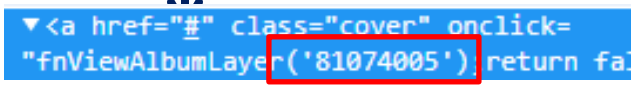
음원 차트 시각화

시각화

■ 크롤링 원리

url : https://www.genie.co.kr/chart/top200





url : https://www.genie.co.kr/detail/albumInfo?axnm=81074005

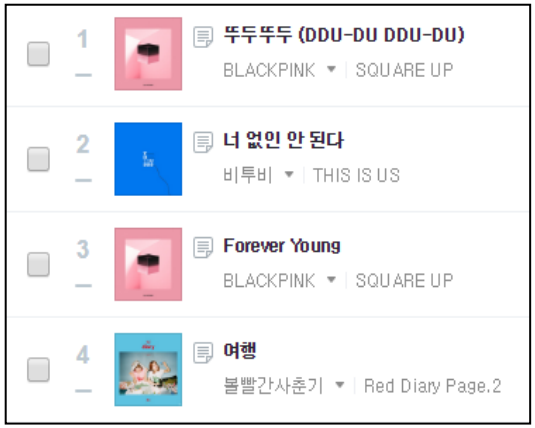
음원 차트 시각화

시각화

■ 크롤링 원리

R

```
for (i in 1: 50) {
u1<-url(chart/top200)
…
//*[@id="body-
content"]/div[6]/div/tab
le/tbody/tr[i]/td[5]/a[1]
…

u2<- url(…
albumInfo?axnm=앨범번호)
}
```



순위, 제목, 가수명, **앨범번호**

음원 차트 시각화

시각화

■ 크롤링 소스

```
top_100 <- function(n){

  chart<-data.frame()
  for (j in 1:2){  #1,2 페이지
    url = paste("http://www.genie.co.kr/chart/top200?ditc=D&ymd=20180614&hh=17&rtm=Y&pg=",j)
    url<-gsub(' ','',url)
    SOURCE = getURL(url)   # url 정보를 가져오고
    PARSED = htmlParse(SOURCE, encoding = "UTF-8")
    print(paste('url:',url))
    for(i in 1:50){
      if(i==1&j==1)
        chart<-get_inf(i,j,PARSED)
      else
        chart<-rbind(chart,get_inf(i,j,PARSED))
    }
  }

  colnames(chart)<-c("순위","제목","아티스트","장르","기획사",'발매일')
```

음원 차트 시각화

시각화

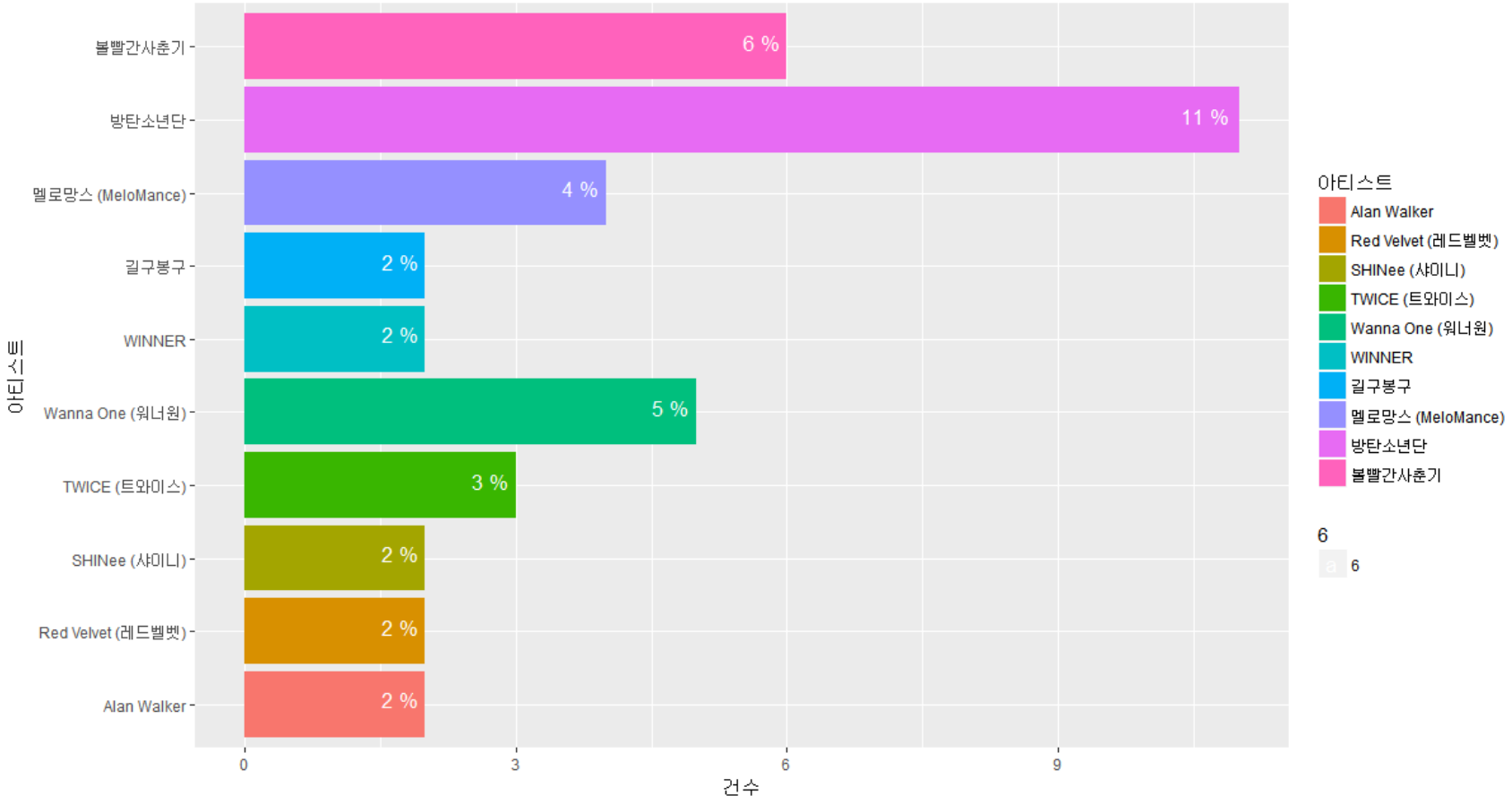
■ 크롤링 소스

```
get_inf<-function(i,j,PARSED){
  album_n<- xpathSApply(PARSED, paste("//*[@id='body-content']/div[6]/div/table/tbody/tr[",i,"]/td[3]/a"), node_onclick)
  album_n<-str_replace_all(album_n,"[^\\d]","" )   # 숫자가 아닌 문자를 지운다.

  inf_url<-gsub(' ','',paste("http://www.genie.co.kr/detail/albumInfo?axnm=",album_n))
  SOURCE2 = getURL(inf_url)   # url 정보를 가져옴
  PARSED_a = htmlParse(SOURCE2, encoding = "UTF-8")

  genre <- xpathSApply(PARSED_a,paste("//*[@id='body-content']/div[2]/div[2]/ul/li[2]/span[2]"),xmlValue) #장르
  agency <- xpathSApply(PARSED_a,paste("//*[@id='body-content']/div[2]/div[2]/ul/li[4]/span[2]"),xmlValue) #기획사
  c_day<- xpathSApply(PARSED_a,paste("//*[@id='body-content']/div[2]/div[2]/ul/li[5]/span[2]"),xmlValue)
  c_day<- gsub('\\.','-',c_day)
  c_day<- gsub('\n','',c_day)


  if (j==2){
    rank<-i+50
  }
  else
    rank<-i
  s_name <-xpathSApply(PARSED,paste("//*[@id='body-content']/div[6]/div/table/tbody/tr[",i,"]/td[5]/a[1]/text()"),xmlValue)
  s_name<-gsub("\n",'',s_name)
  s_name<-gsub(' ','',s_name)
  artist <-xpathSApply(PARSED,paste("//*[@id='body-content']/div[6]/div/table/tbody/tr[",i,"]/td[5]/a[2]"),xmlValue)

  return(c(rank,s_name,artist,genre,agency,c_day))
}
```

음원 차트 시각화

시각화

1. 최근 인기 있는 가수는 누구일까?

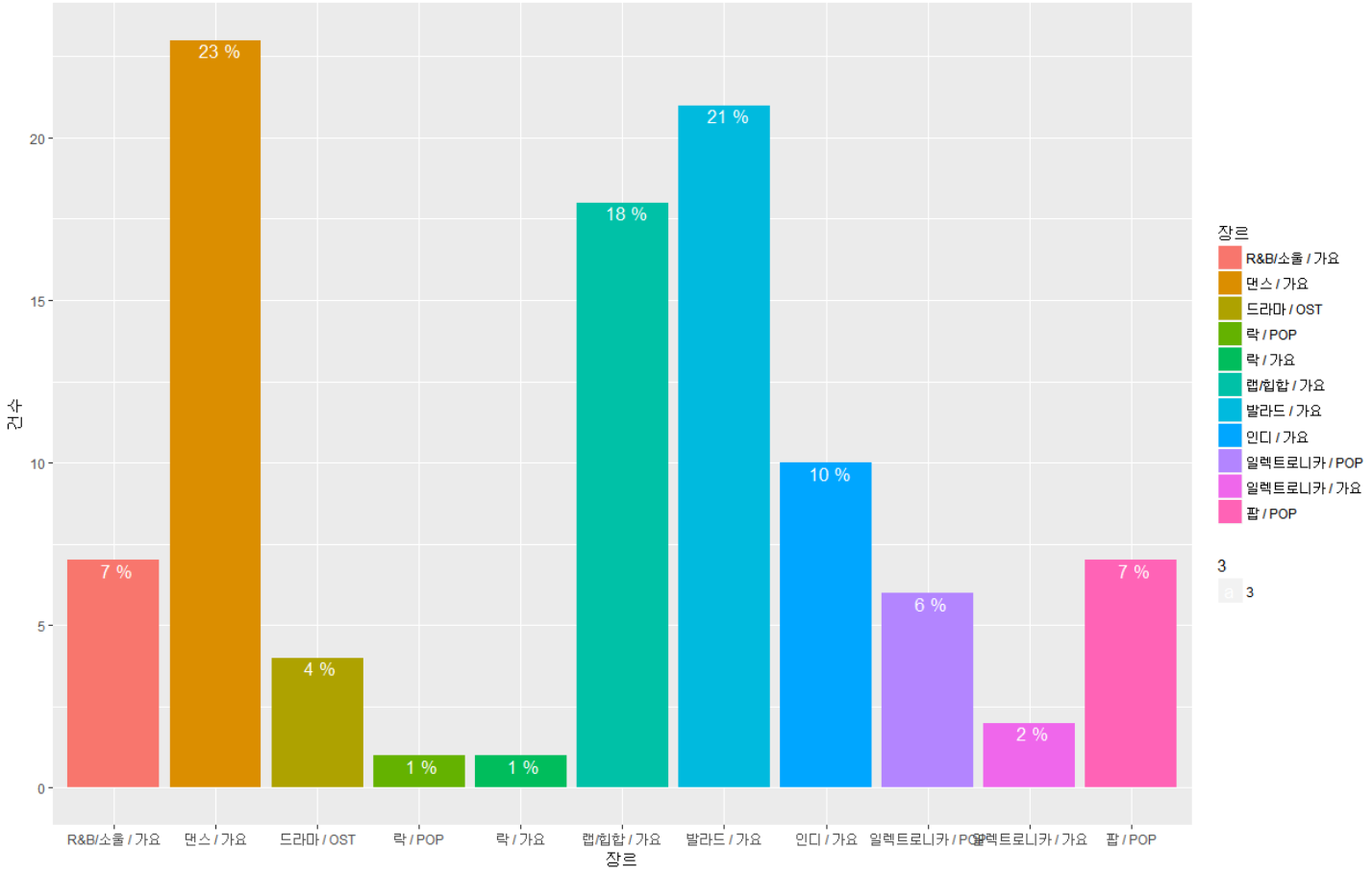


현재 지니 차트 top100에서 가수별 노래 점유율

음원 차트 시각화

시각화

2. 최근 유행하는 음악 장르는 무엇인가?

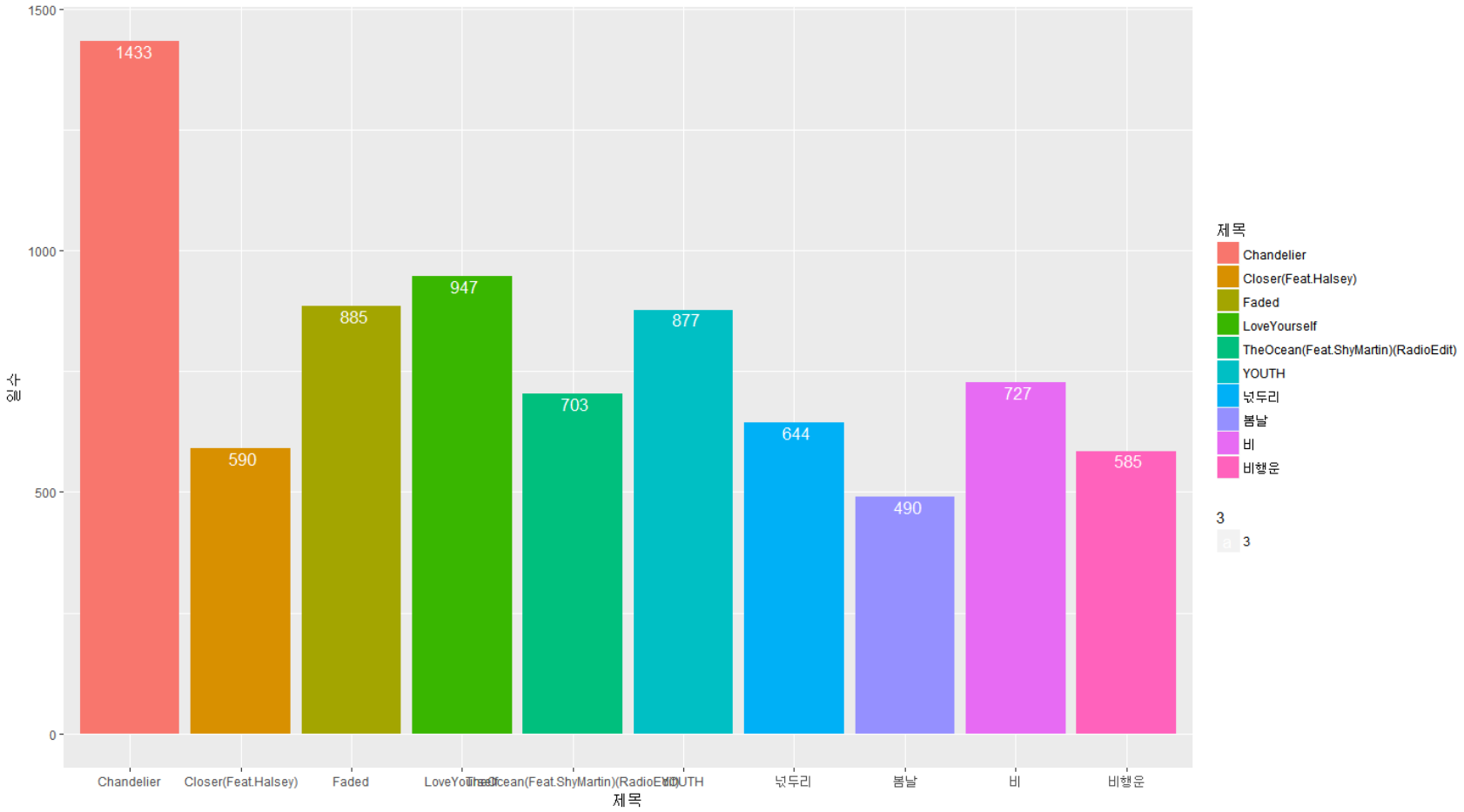


현재 지니 차트 top100에서 장르별 노래 점유율

음원 차트 시각화

시각화

3. 오래된 노래가 인기가 있을까?

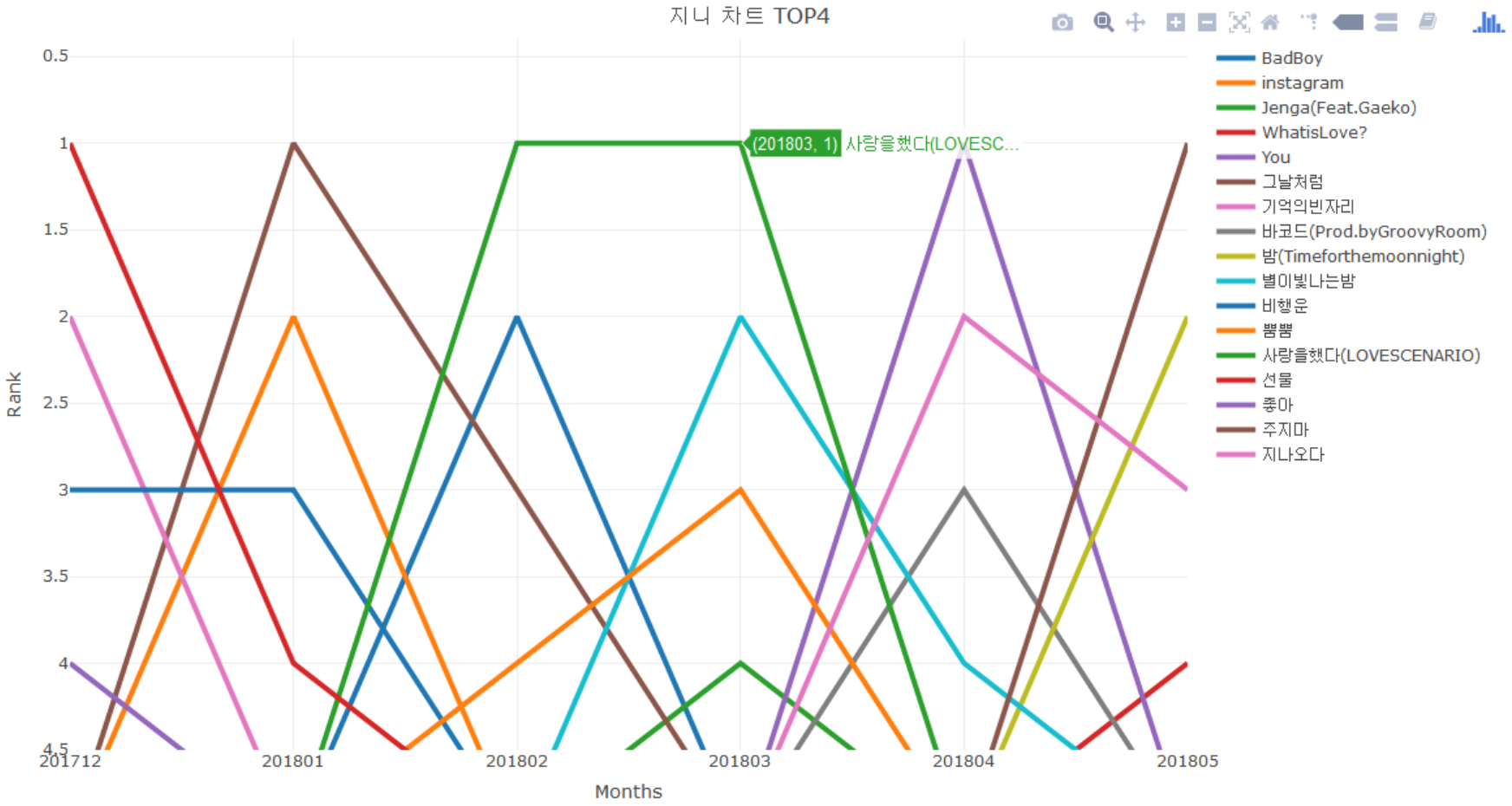


현재 지니 차트 top100에서 발매일이 가장 오래 전인 노래

음원 차트 시각화

시각화

4. 히트곡이 얼마나 오랫동안 최상위권을 유지 할 수 있을까?

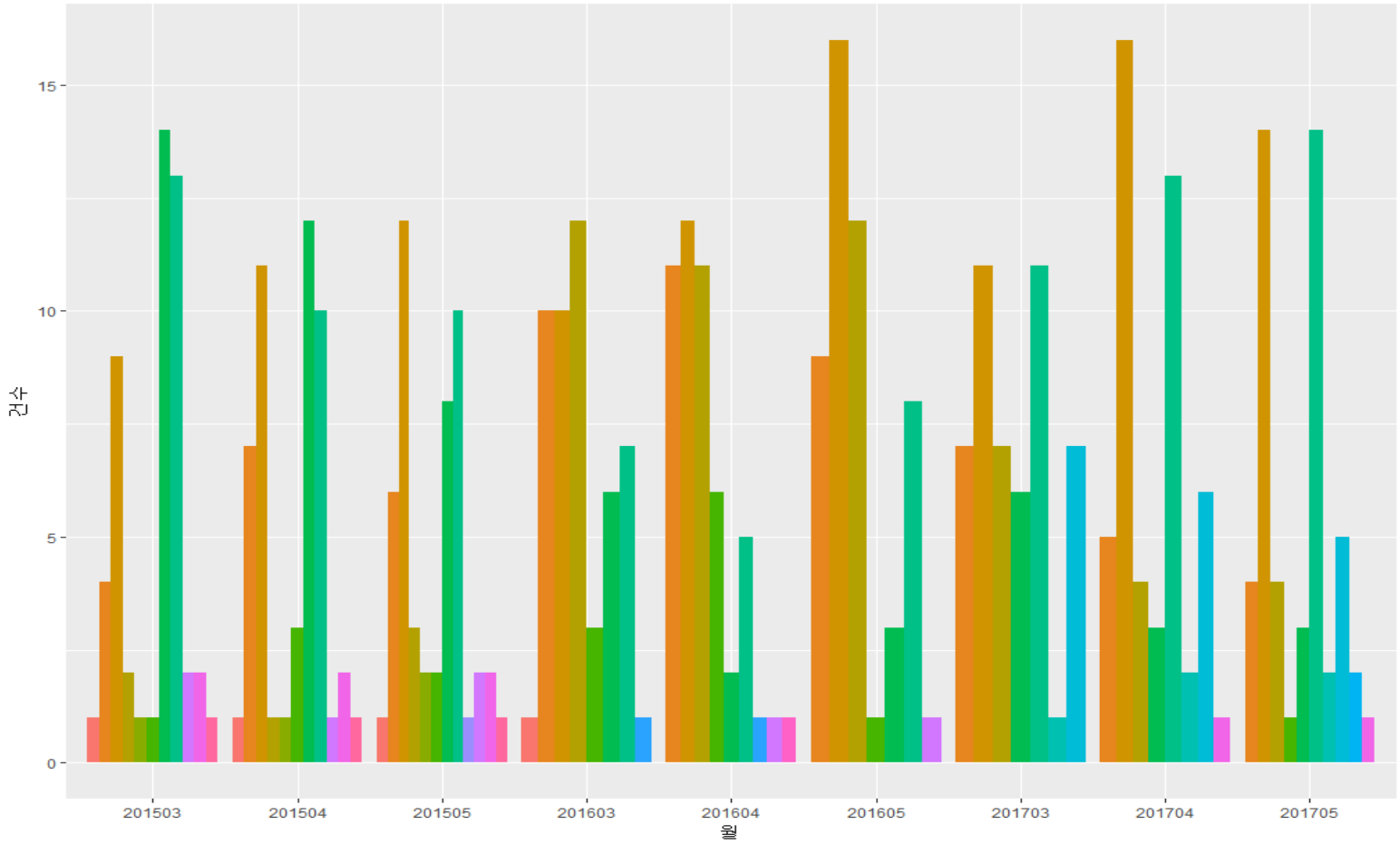


현재 지니 차트 top100에서 가수별 노래 점유율

음원 차트 시각화

시각화

5. 계절이 음악장르 인기에 영향을 미칠까? (봄)



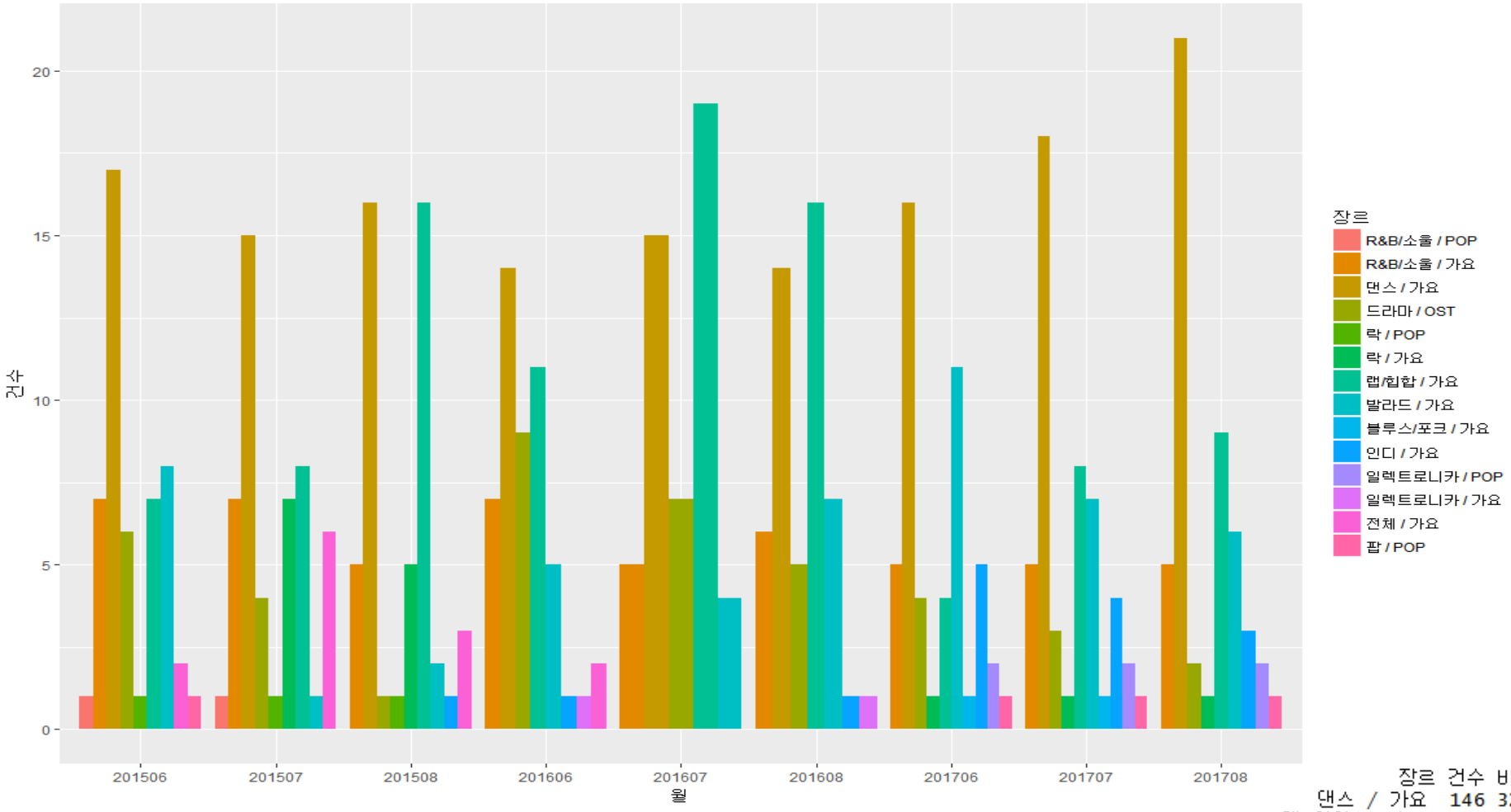
장르	건수	비율
댄스 / 가요	111	24.7
발라드 / 가요	91	20.2
R&B/소울 / 가요	63	14.0
랩/힙합 / 가요	57	12.7
드라마 / OST	56	12.4
인디 / 가요	18	4.0
락 / 가요	17	3.8
팝 / POP	8	1.8

최근 3년 3,4,5월 top 50에서의 장르 점유율

음원 차트 시각화

시각화

5. 계절이 음악장르 인기에 영향을 미칠까? (여름)



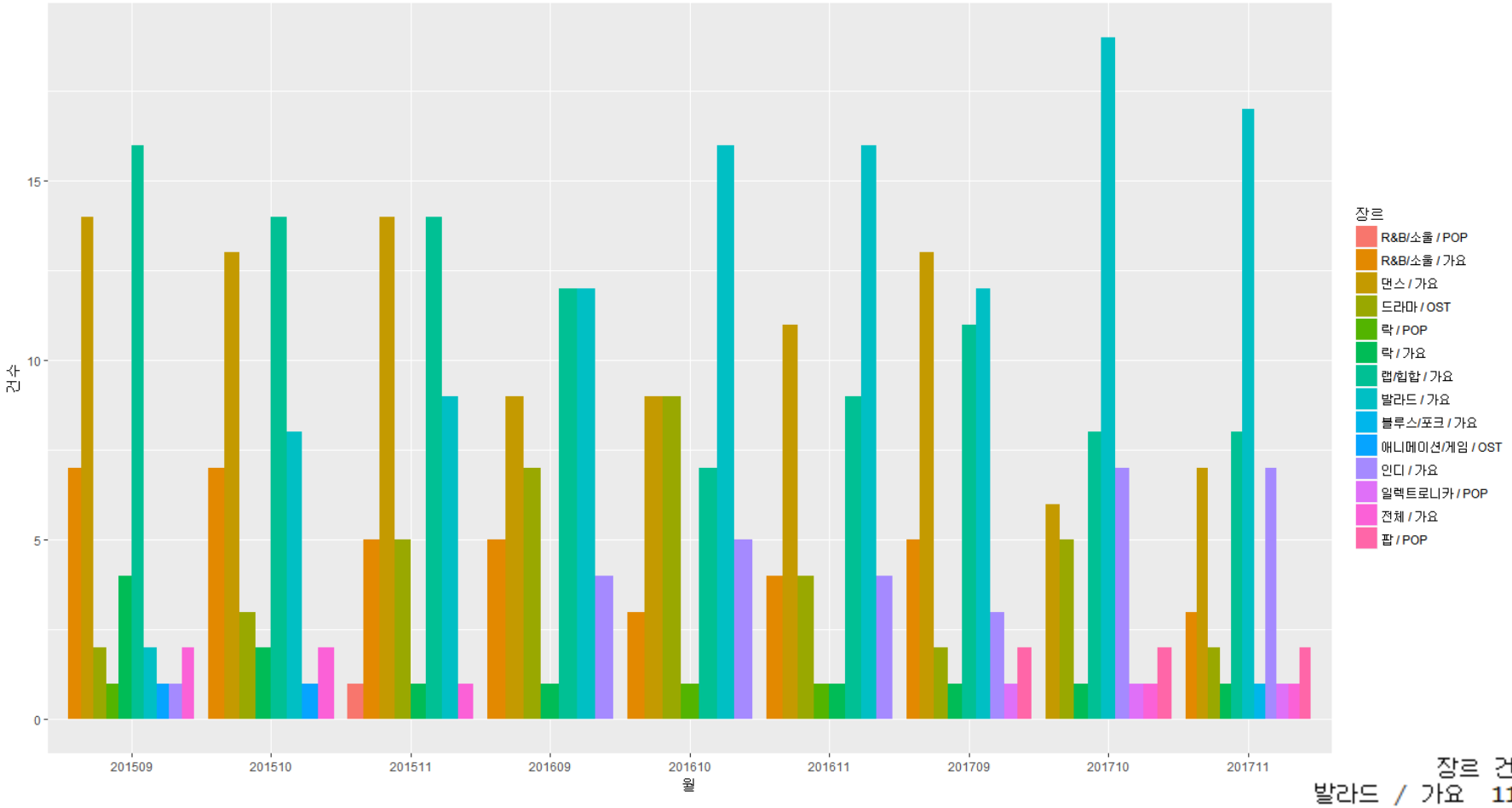
장르	건수	비율
댄스 / 가요	146	32.4
랩/힙합 / 가요	98	21.8
R&B/소울 / 가요	52	11.6
발라드 / 가요	51	11.3
드라마 / OST	41	9.1
락 / 가요	15	3.3
인디 / 가요	15	3.3

최근 3년간 6,7,8월 top 50에서의 장르 점유율

음원 차트 시각화

시각화

5. 계절이 음악장르 인기에 영향을 미칠까? (가을)



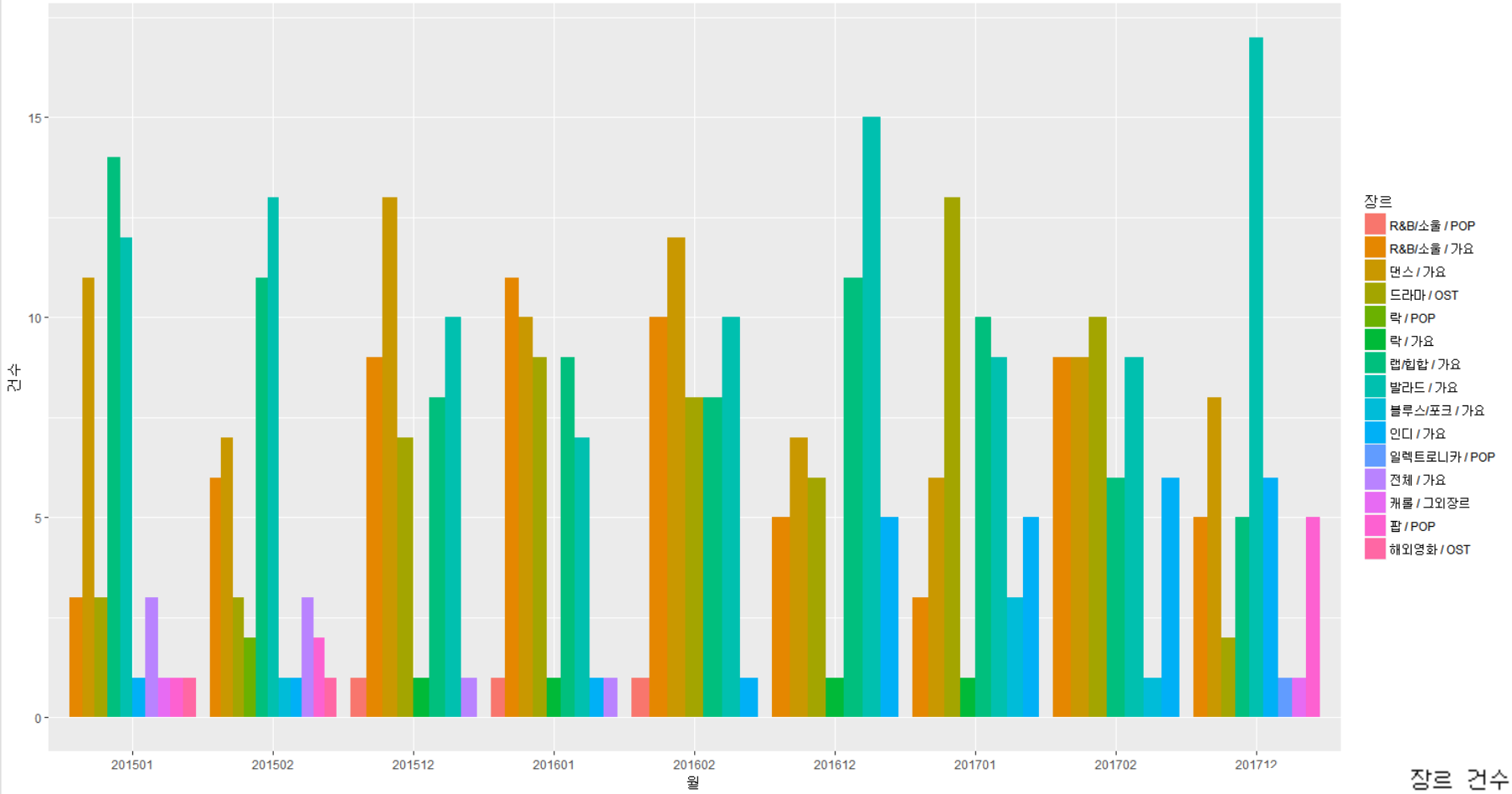
장르	건수	비율
발라드 / 가요	111	24.7
랩/힙합 / 가요	99	22.0
댄스 / 가요	96	21.3
R&B/소울 / 가요	39	8.7
드라마 / OST	39	8.7
인디 / 가요	31	6.9
락 / 가요	12	2.7

최근 3년 9,10,11월 top 50에서의 장르 점유율

음원 차트 시각화

시각화

5. 계절이 음악장르 인기에 영향을 미칠까? (겨울)



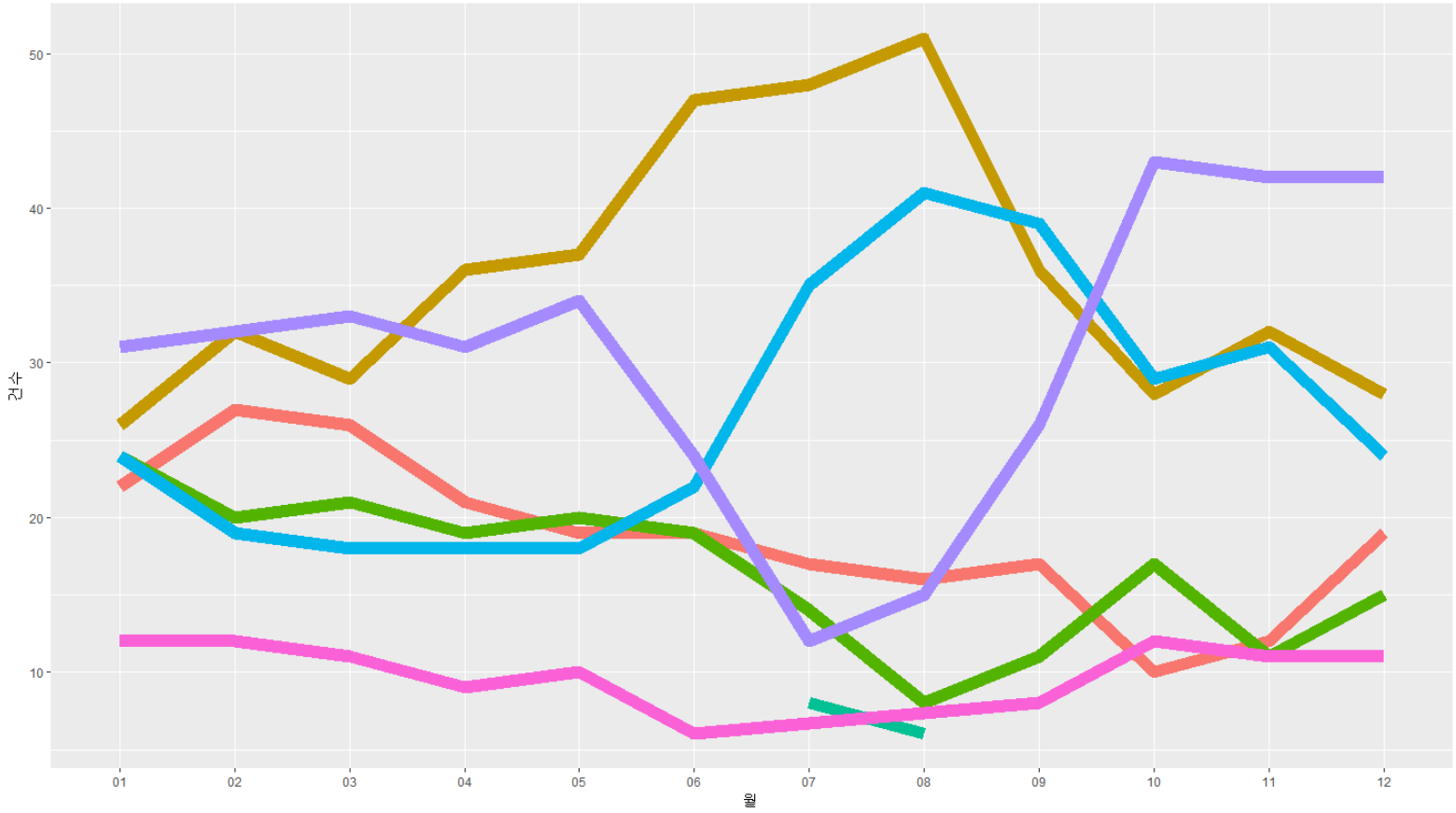
장르	건수	비율
발라드 / 가요	102	22.7
댄스 / 가요	83	18.4
랩/힙합 / 가요	82	18.2
R&B/소울 / 가요	61	13.6
드라마 / OST	61	13.6
인디 / 가요	26	5.8

최근 3년 1,2,12월 top 50에서의 장르 점유율

음원 차트 시각화

시각화

5. 계절이 음악장르 인기에 영향을 미칠까? (최근 3년 전체)



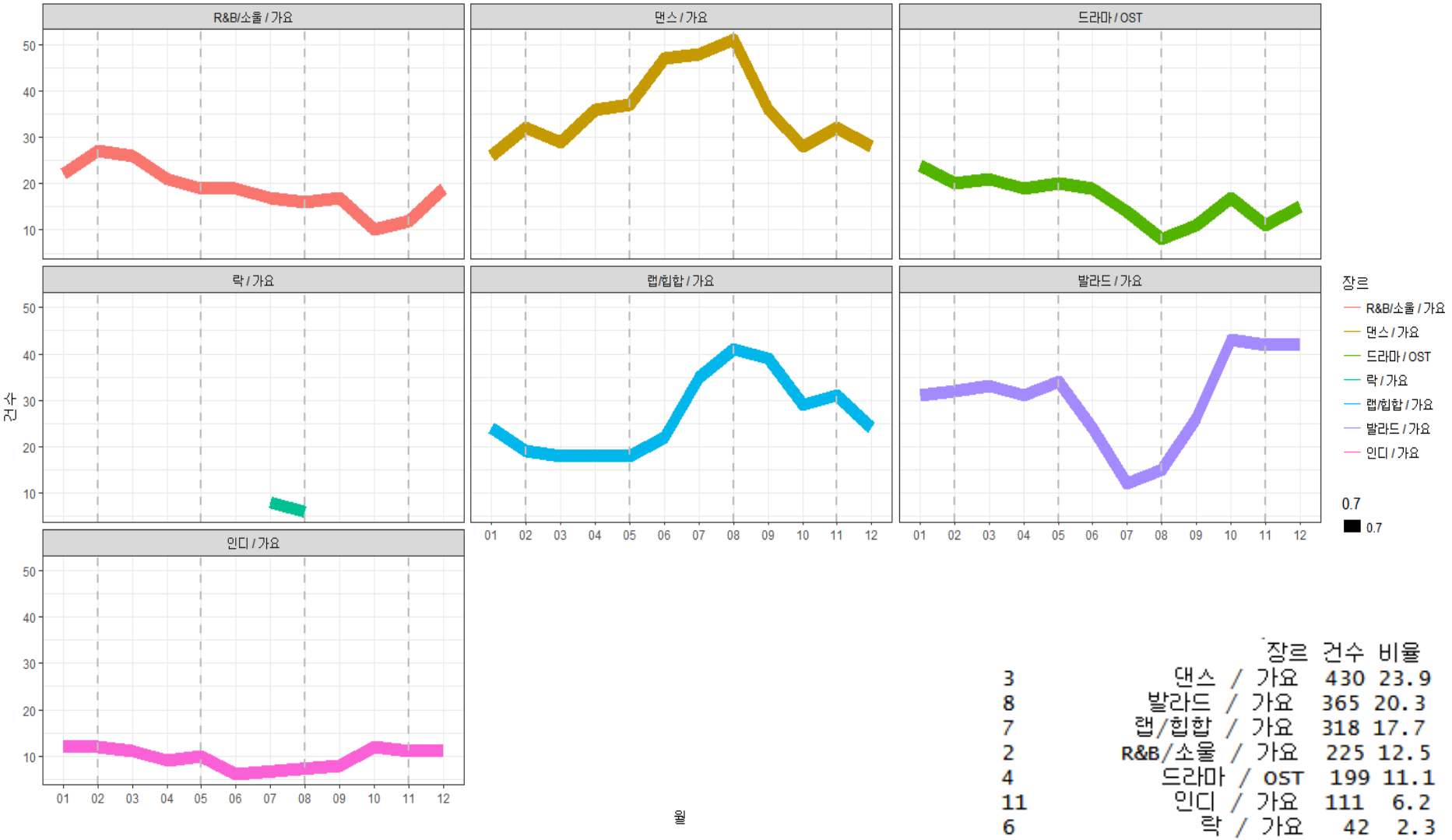
	장르	건수	비율
3	댄스 / 가요	430	23.9
8	발라드 / 가요	365	20.3
7	랩/힙합 / 가요	318	17.7
2	R&B/소울 / 가요	225	12.5
4	드라마 / OST	199	11.1
11	인디 / 가요	111	6.2
6	락 / 가요	42	2.3
16	팝 / POP	35	1.9

최근 3년간 top 50의 월별 장르 건수

음원 차트 시각화

시각화

5. 계절이 음악장르 인기에 영향을 미칠까? (최근 3년 월별 장르 인기 추이)

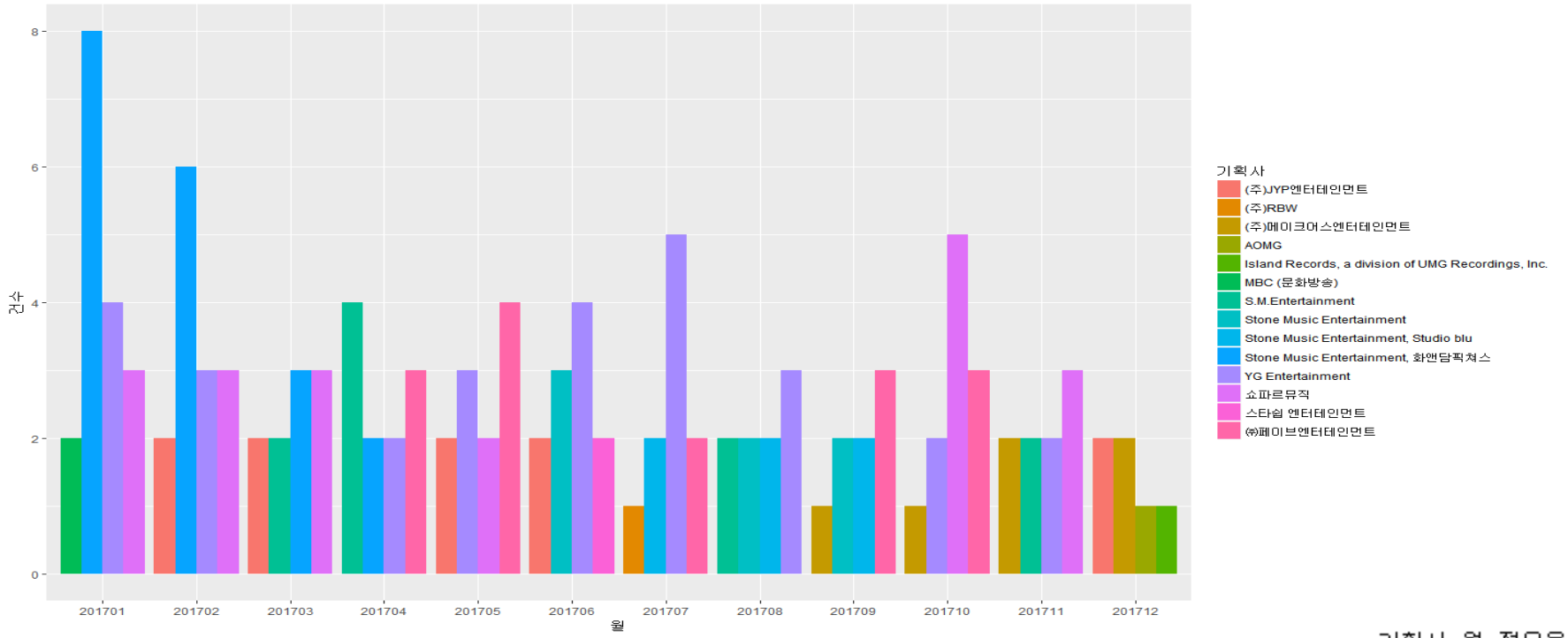


	장르	건수	비율
3	댄스 / 가요	430	23.9
8	발라드 / 가요	365	20.3
7	랩/힙합 / 가요	318	17.7
2	R&B/소울 / 가요	225	12.5
4	드라마 / OST	199	11.1
11	인디 / 가요	111	6.2
6	락 / 가요	42	2.3
16	팝 / POP	35	1.9

음원 차트 시각화

시각화

6. 작년엔 어떤 기획사의 가수들이 활약했을까?



작년 차트의 top50에서 기획사별 노래 건수

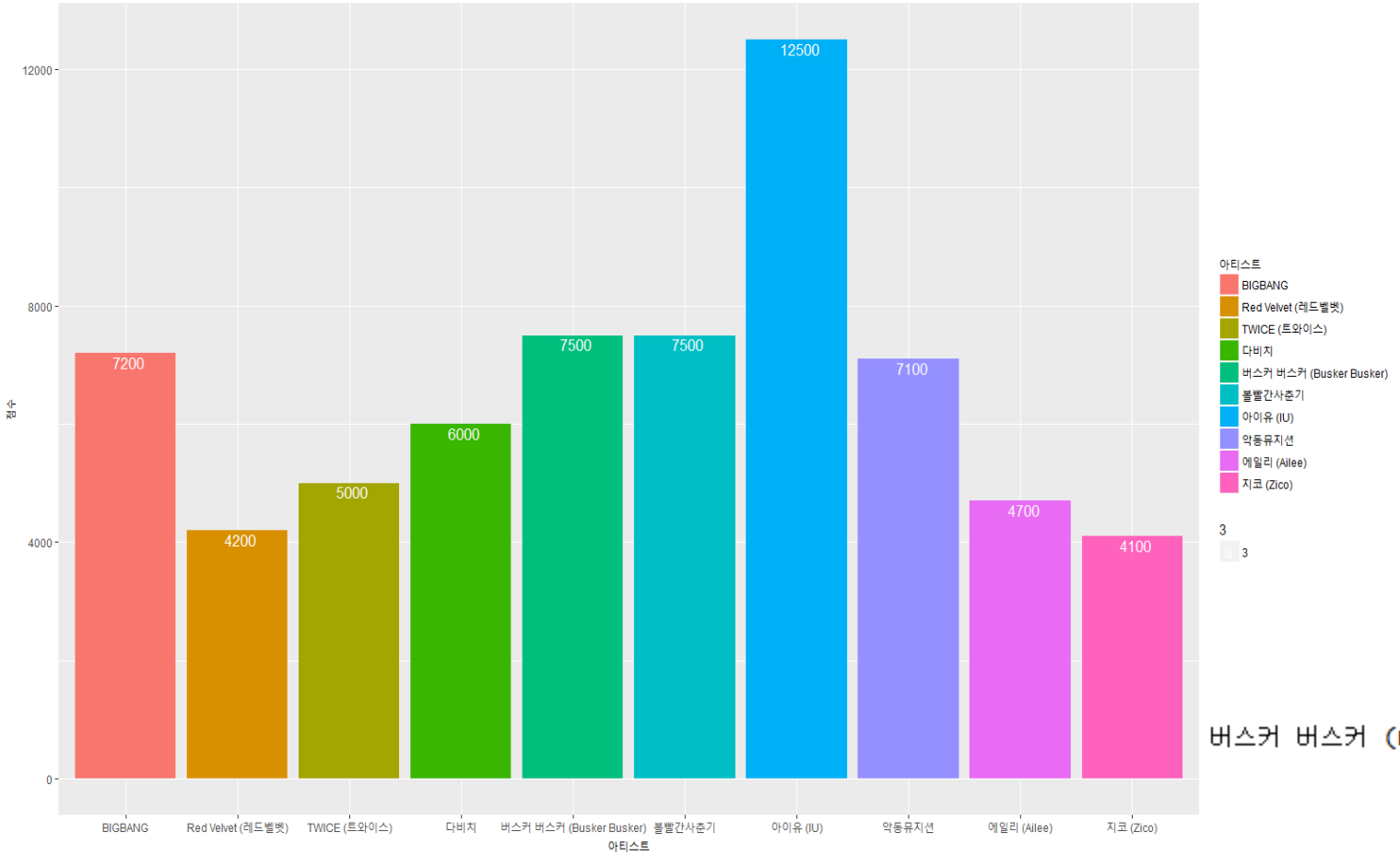
기획사	월	점유율
YG Entertainment	32	13.3
쇼파르뮤직	26	10.8
Stone Music Entertainment, 화앤담픽쳐스	21	8.8
㈜페이브엔터테인먼트	19	7.9
S.M.Entertainment	14	5.8
(주)JYP엔터테인먼트	13	5.4
Stone Music Entertainment	8	3.3
Stone Music Entertainment, Studio blu	8	3.3
(주)메이크어스엔터테인먼트	7	2.9
Atlantic Records UK	5	2.1
미스틱엔터테인먼트	5	2.1
세븐시즌스	5	2.1
스타쉽 엔터테인먼트	5	2.1

전체 기획사 수 = 51

음원 차트 시각화

시각화

7. 2012.2.01~ 현재 (76개월) 사이에 인기가 많은 가수는?



아티스트	점수
아이유 (IU)	12500
버스커 버스커 (Busker Busker)	7500
볼빨간사춘기	7500
BIGBANG	7200
악동뮤지션	7100
다비치	6000
TWICE (트와이스)	5000
에일리 (Ailee)	4700
Red Velvet (레드벨벳)	4200
지코 (Zico)	4100

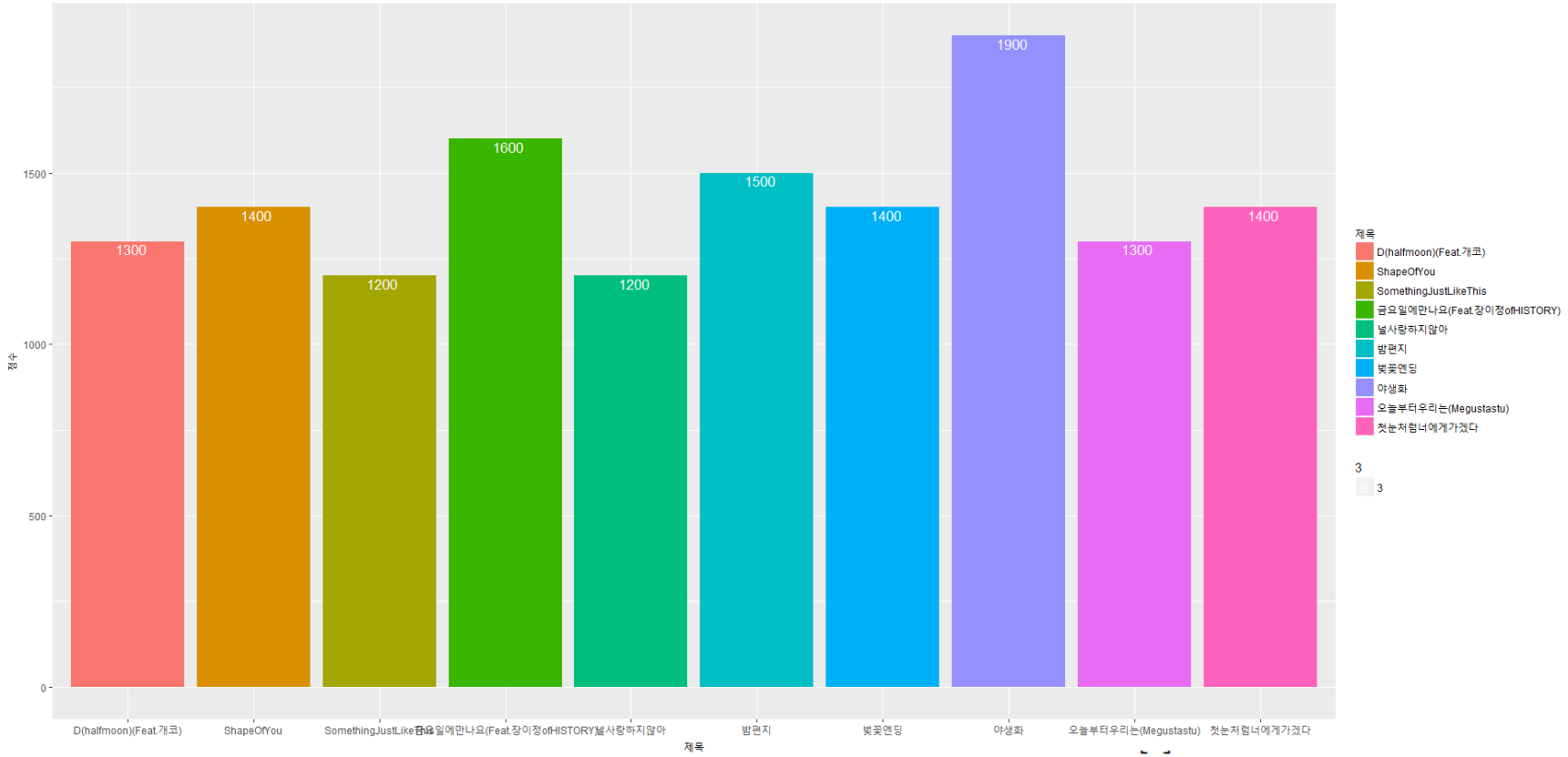
2012년 2월 차트부터 top 50 안에 드는 노래가 있을 경우 +100점

(소스는 [101점 –등수] 로 수정 ex. 5등일 경우 101 – 5 = 96점)

음원 차트 시각화

시각화

8. 2012.2.01~ 현재 (76개월) 사이의 히트곡은?



2012년 2월 차트부터 top 50 안에 드는 노래가 있을 경우 +100점

(소스는 [101점 –등수] 로 수정 ex. 5등일 경우 101 – 5 = 96점)

	제목	점수
1201	야생화	1900
635	금요일에만나요(Feat.장이정ofHISTORY)	1600
971	밤편지	1500
422	ShapeOfYou	1400
975	벚꽃엔딩	1400
1444	첫눈처럼너에게가겠다	1400
138	D(halfmoon)(Feat.개코)	1300
1266	오늘부터우리는(Megustastu)	1300
432	SomethingJustLikeThis	1200
779	널사랑하지않아	1200

결론

시각화

1. 최근 인기 있는 가수는 누구일까?

최근 인기가 많은 가수는 방탄소년단, 빨간 사춘기, 워너원, 멜로망스 등이 있으며, 특히 2자리수 점유율을 보이는 **방탄소년단**의 인기가 두드러진다.

2. 최근 인기 유행하는 음악 장르는 무엇인가?

1-100위의 노래 중 댄스/가요, 발라드/가요, 랩/힙합 장르는 각각 20% 정도의 점유율을 보이며 락 장르는 2%의 점유율로 낮은 비중을 차지한다.

3. 오래된 노래가 인기가 있을까?

top100의 차트에서 발매한지 가장 오래된 노래는 2014년 7월에 발매한 sia의 chandelier이다.
가장 오래된 노래 1-5위 까지 모두 외국 노래이고, 오래된 한국 노래로는 폴킴의 비, 닐로의 넋두리, 문문의 비행운이 있다.

4. 히트곡이 얼마나 오랫동안 최상위권을 유지 할 수 있을까?

최근 6개월 동안 2달 연속 4위 이내의 순위를 기록한 노래는 선물 (멜로망스, 1~4) , 사랑을 했다 (iKON , 1~1) , 별이 빛나는 밤 (마마무, 2~4) 위를 기록했고, 한번 떨어진 순위가 다시 올라가는 경우는 없었다.

결론

시각화

5. 계절이 음악장르 인기에 영향을 미칠까? (최근 3년 전체)

봄 (댄스 24%,발라드 20%), 여름(댄스 32%, 랩/힙합 21%), 가을 (발라드 24%, 랩/힙합 22%), 겨울 댄스 (발라드 22%, 댄스 18%)의 점유율을 보이며 특히 댄스, 발라드가 항상 40%이상의 점유율을 보인다. 봄-여름은 댄스장르가 강세를 보이며 특히, 여름은 댄스 장르가 30%이상의 높은 점유율을 보인다. 가을-겨울은 발라드가 강세를 보이며 겨울은 댄스장르가 10%대로 하락하지만 2위의 점유율을 보인다.

6. 작년엔 어떤 기획사의 가수들이 활약했을까?

작년 top100에서 YG가 13%, 쇼파르뮤직 10%, 화앤담픽쳐스&Stone Music Entertainment 8%의 점유율을 차지했으며, SM과 JYP는 5%의 점유율을 기록했다.

7. 2012.2.01~ 현재 (76개월) 사이에 인기 가수와 히트곡은?

아이유(12000), 볼빨간사춘기, 버스커버스커(7500), 빅뱅(7200) 이 꾸준히 인기가 있었고,
야생화(1900),밤편지(1600), 금요일에 만나요(1500) 등의 노래가 가장 높은 점수를 기록했다.

감사합니다.