

The background is a light pink color. In the top left, there is a large, light blue abstract shape. In the top right, there is a large, light red abstract shape. In the bottom left, there is a large, light blue abstract shape. In the bottom right, there is a large, light orange abstract shape. There are also some small, light blue and light orange dots scattered around. On the left side, there are three white wavy lines. On the right side, there is a cluster of small black dots.

빅데이터 기반 프로젝트

들어볼래?

1조 김유미, 김화영, 최가람  
연세 빅데이터2023

# 💡 목차

## 1. 프로젝트 개요

- 구성원/일정/사용도구 소개
- 프로젝트 기획 배경

## 2. 데이터 구성 및 활용

- 데이터 수집(크롤링)
- 데이터 시각화

## 3. 데이터 처리 및 개발 과정

- 추천 시스템
- 웹 개인화

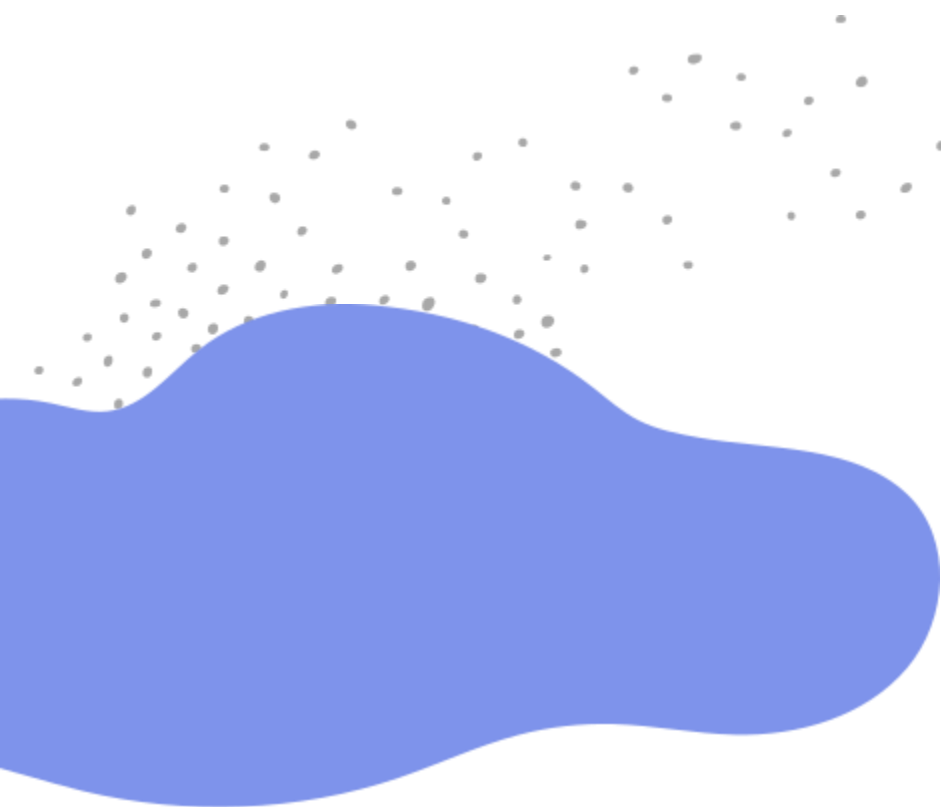
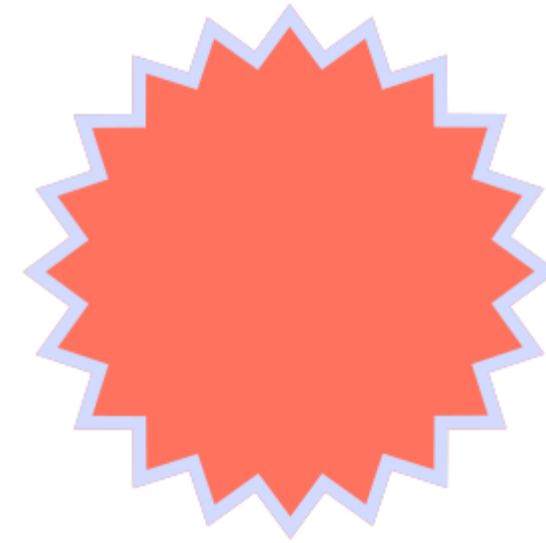
## 4. 프로젝트 리뷰

- 기대효과
- 가능성 및 한계
- 소감



# 1. 프로젝트 개요

- 구성원/일정/사용 도구 소개
- 프로젝트 기획 배경



# '도전했조!' 의 구성원 소개/역할



김유미 (조장)

시스템 설계  
명곡 추천  
웹페이지 구현, 크롤링



김화영

발표  
플레이리스트 추천  
프로젝트 관리, 크롤링



최가람

Dataset 가공 및 분석  
DB, 컨텐츠 추천  
PPT, 크롤링

# 상세일정

## 01

05/24 ~ 05/26

사전계획 및 주제선정  
기획안 작성  
자료조사 및 기획안 발표

## 03

06/04 ~ 06/06

웹 flask 제작  
웹 개인화 작업

## 02

05/30 ~ 06/03

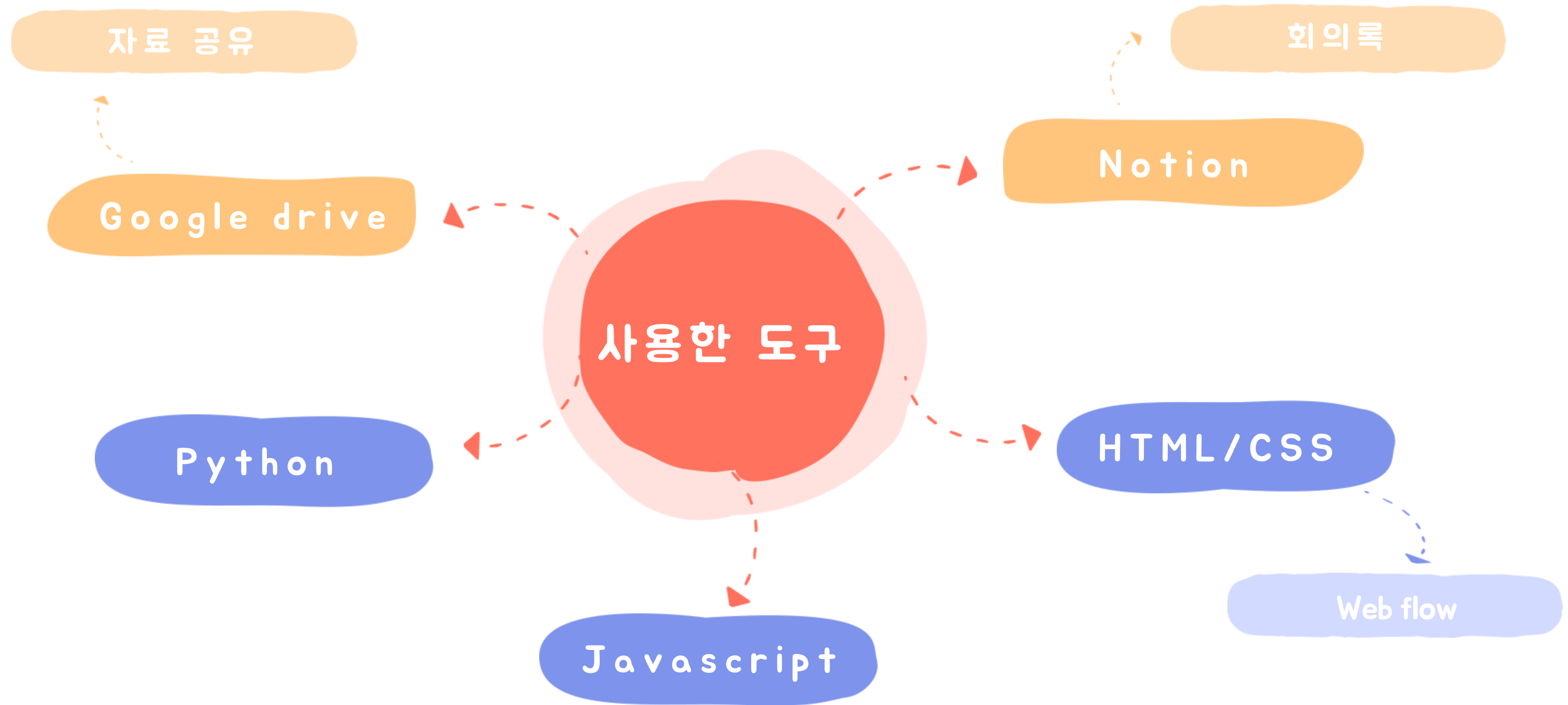
데이터 수집 및 전처리  
추천 시스템 구현  
웹 flask 제작

## 04

06/07 ~ 06/09

테스트  
PPT제작 및 발표

# 프로젝트 관리





# 프로젝트 기획 배경

노래 가사 요약  
시스템 개발  
But 능력의 한계  
(딥러닝 요구)



우리 능력  
구현 할 수 있는  
서비스 필요



빅데이터셋구축  
머신러닝  
추천시스템  
노래 추천을 해볼까?




What?  
사용자가 듣고 싶은  
비슷한  
잘 알려지지 않는  
다른 사람들이 듣는



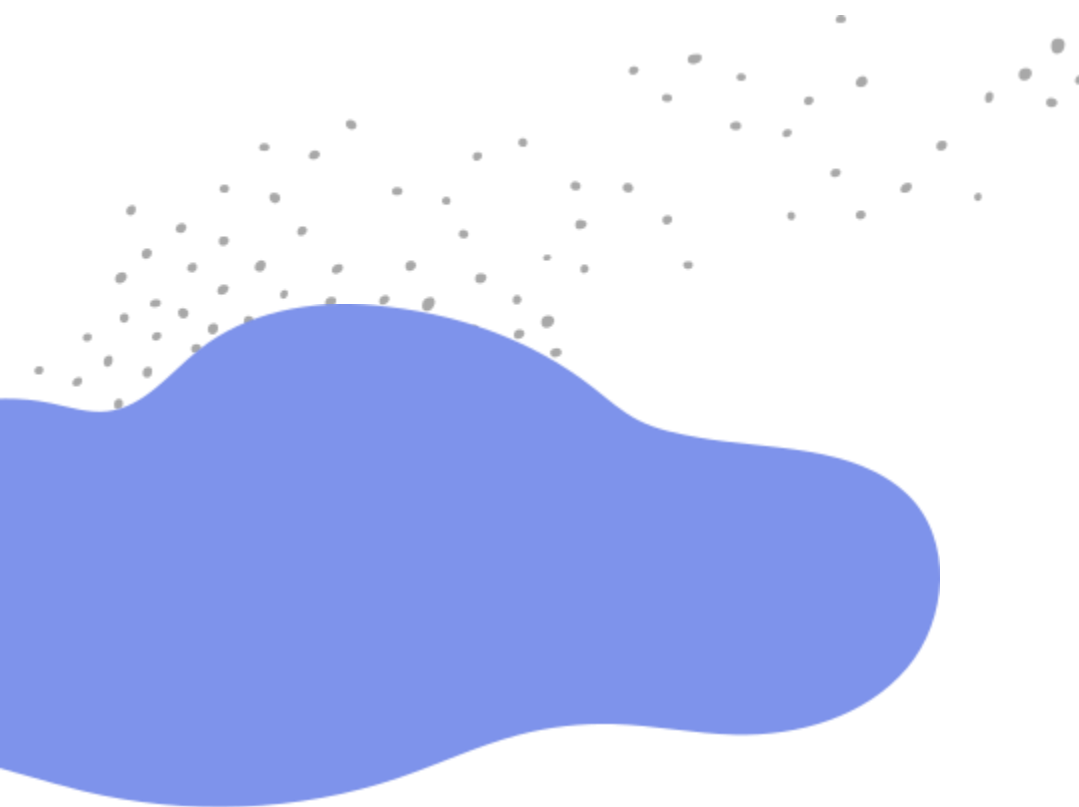
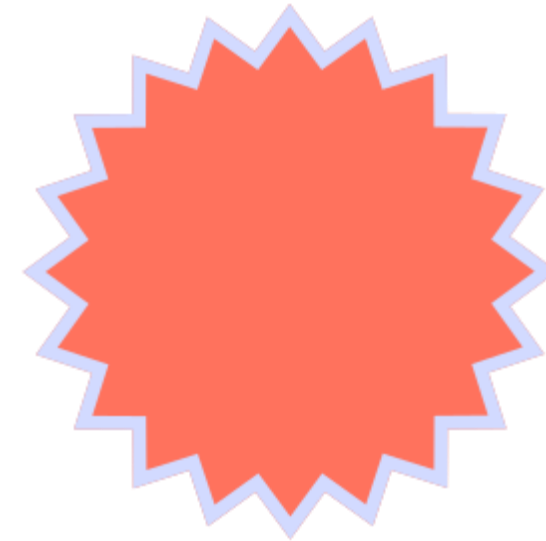
노래 정보 수집  
노래가사 코사인유사도(cosine similarity) 개발  
웹 페이지에 적용  
노래 추천 시스템!





## 2. 데이터 구성 및 활용

- 데이터 수집(크롤링)
- 데이터 시각화





## 2. 데이터 구성 및 활용

# Melon 크롤링

Dc 디지털데일리



실시간  
뉴스

가전

삶을 '비스포크' 하다...삼성전자 '비스포크 라이프 2023' 행사 개최






콘텐츠

## 멜론 월간 이용자 수 677만명...삼성 더하니 시장점유율 절반 육박

디지털데일리 | 발행일 2023-02-09 10:40:15

이나연



1		멜론(Melon) Kakao Entertainment Corp.	6,775,948	29.89%
2		YouTube Music Google LLC	5,053,452	22.29%
3		Samsung Music - 삼성 뮤직 Samsung Electronics Co., Ltd.	4,346,268	19.17%
4		지니뮤직 - genie (주)지니뮤직	3,376,641	14.89%
5		FLO - 플로 Dreamus Company	2,136,225	9.42%

멜론 티켓

 이용권구매 | 멜론라운지 | 이벤트 | 공지사항

Melón

인기가요 미공개 영상보고 최대 사진 받는 법!



급상승

1. 태연 (TAEYEON) +9

멜론차트

최신음악

장르음악

멜론DJ

멜론TV

스타포스트

매거진

뮤직어워드

어학

마이뮤직

TOP100

일간

주간


월간

시대

Q 차트 파인더

TOP100 ?


2023.06.08 15:00 ∨

 셔플듣기

 전체듣기


 들기

 + 담기

 다운

 FLAC

 선물

새로고침 

☐ 순위

곡정보

앨범

좋아요

듣기

담기

다운

유비

☐ 1

—



퀸카 (Queencard)  
(여자)아이들

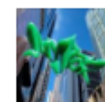
I feel

♡ 83,388



☐ 2

—



Spicy  
aespa

MY WORLD - The 3rd Mi...

♡ 90,764



☐ 3

—



I AM  
IVE (아이브)

I've IVE

♡ 165,630



☐ 4

—



UNFORGIVEN (feat. Nile Rod...  
LE SSERAFIM (르세라핌)

UNFORGIVEN

♡ 96,478



☐ 5

—



Kitsch  
IVE (아이브)

I've IVE

♡ 134,613



☐ 6

—



모래 알갱이  
임영웅

모래 알갱이

♡ 33,472



# 데이터 수집

## "연도별"

1964년 ~ 2022년  
연도별 인기순  
최대 50곡

Melón [시대별 차트] 연도별로 보는 그 시절 유행곡! Q 급상승 4. 인생아 고미쳤다 NEW 대한민국 대표 보컬리스트 소향 새 싱글 [EXODUS]

멜론차트 최신음악 장르음악 멜론DJ 멜론TV 스타포스트 매거진 뮤직어워드 어학 마이뮤직

TOP100 일간 주간 월간 시대 Q 차트 파인더

시대별 차트 ?

국내
해외

< 1964년도 > 1960년대 1964년

## "장르별"

발라드, 댄스, 랩/힙합,  
R&B/Soul, 인디음악, 록/메탈,  
트로트, 포크/블루스,  
국내영화/드라마/뮤지컬  
인기순 최대 500곡

Melón 새로운 서사의 시작! 감디너벨 [Wasteland] Q 급상승 4. 인생아 고미쳤다 NEW 대한민국 대표 보컬리스트 소향 새 싱글 [EXODUS]

멜론차트 최신음악 장르음악 멜론DJ 멜론TV 스타포스트 매거진 뮤직어워드 어학 마이뮤직

한국대중음악 해외POP음악 그외인기장르

발라드	댄스	랩/힙합	R&B/Soul	인디음악	록/메탈	트로트
포크/블루스						

## "플레이리스트"

인기 테마당 200개  
기분전환, 비오는 날, 드라이브,  
카페, 한강, 버스 등

Melón [음악의 차트] 밤눈에 뜨는 주간 차트 리뷰 Q 급상승 1. 송시경 + 54 대한민국 대표 보컬리스트 소향 새 싱글 [EXODUS]

멜론차트 최신음악 장르음악 멜론DJ 멜론TV 스타포스트 매거진 뮤직어워드 어학 마이뮤직

투데이 전문가 선곡 #테마장르 파워DJ 인기

#테마장르 인기 테마/장르 태그 # 원하는 태그를 입력해주세요. Q DJ 신청하기

인기 테마
인기 장르

기분전환	감성	힐링	드라이브	사랑	추억	이별
여행	여름	휴식	운동	비오는날	분위기	위로
트렌디	공부	몽환	ASMR	카페	클럽	매장
노래방	버스	라운지	한강	집	지하철	

# 총 데이터 갯수



총 6243 곡

추천시스템에 활용할 분석 데이터



Play list  
3845 개

대중적인 테마 분석 데이터

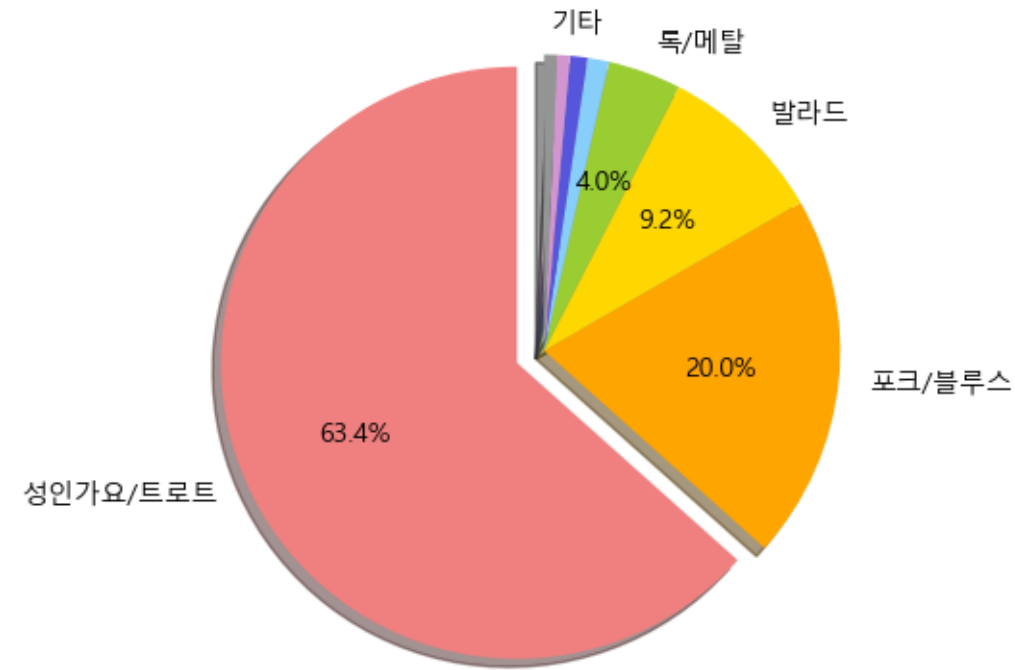


Play list내의 곡  
60271 개

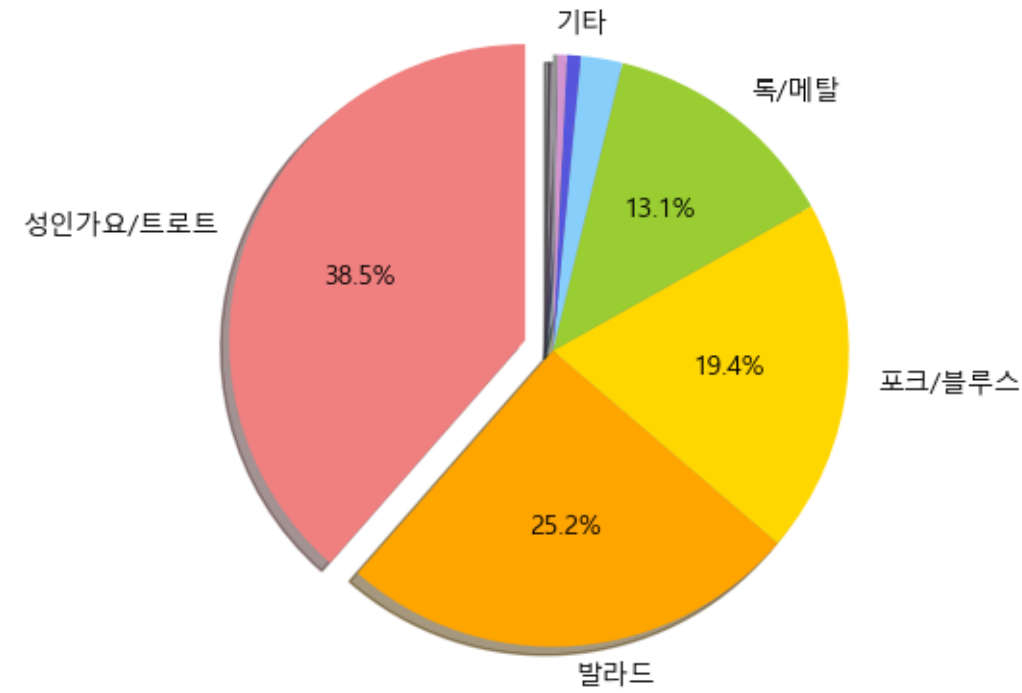
웹 구현 데이터

# 01 연도별 장르 분포도

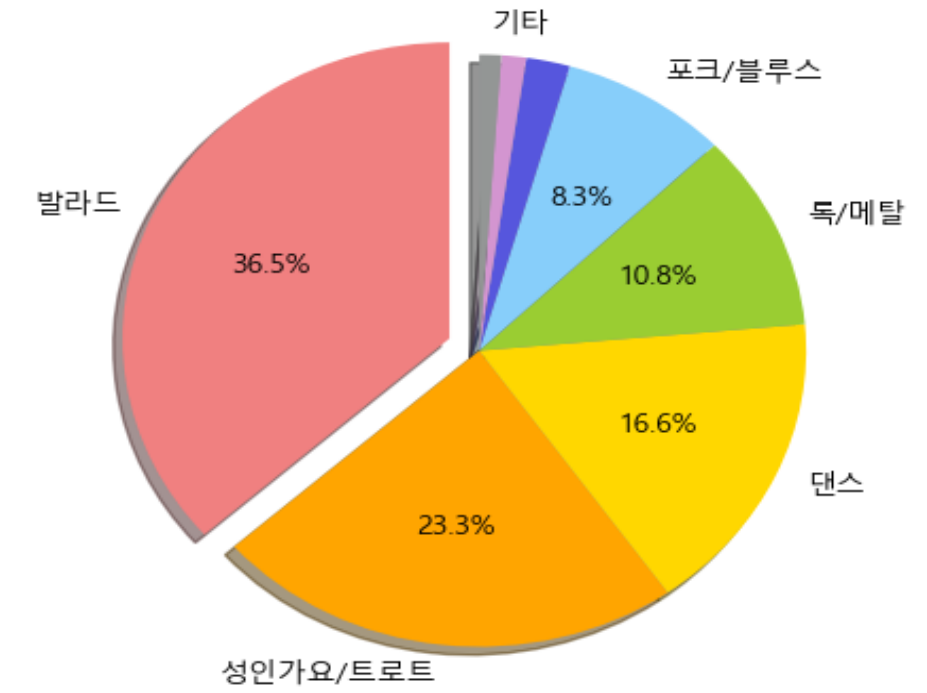
1960~70년 장르분포



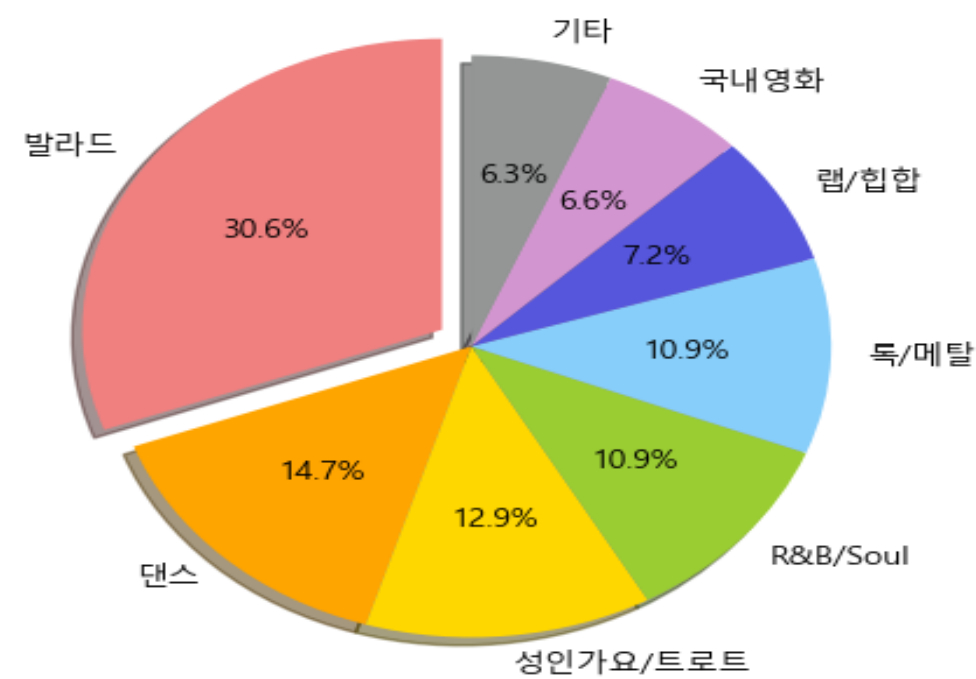
1980년 장르분포



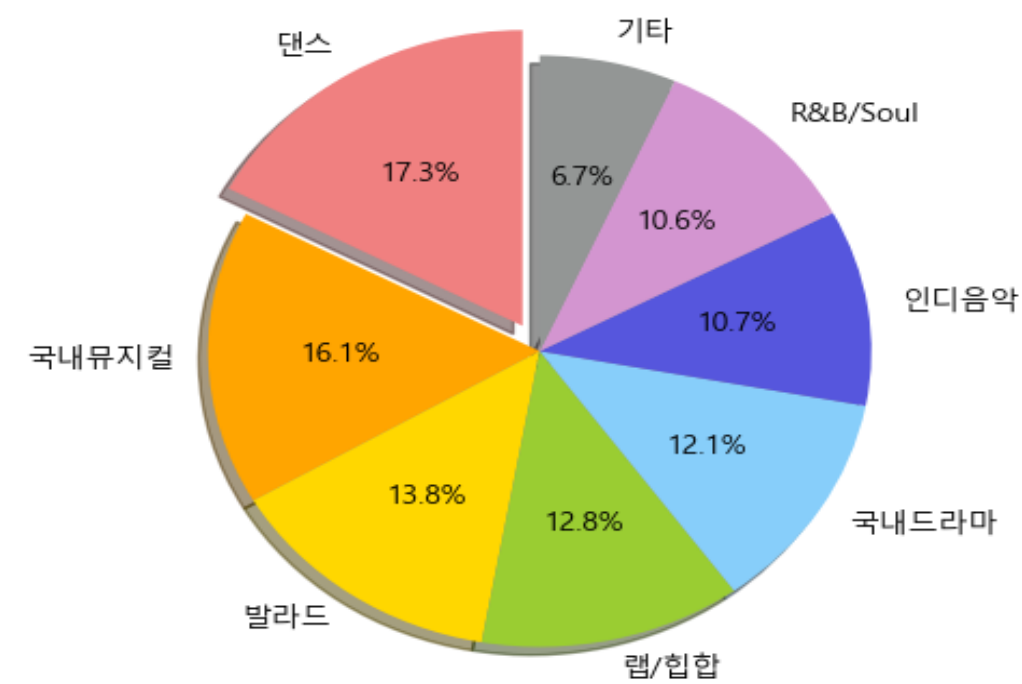
1990년 장르분포



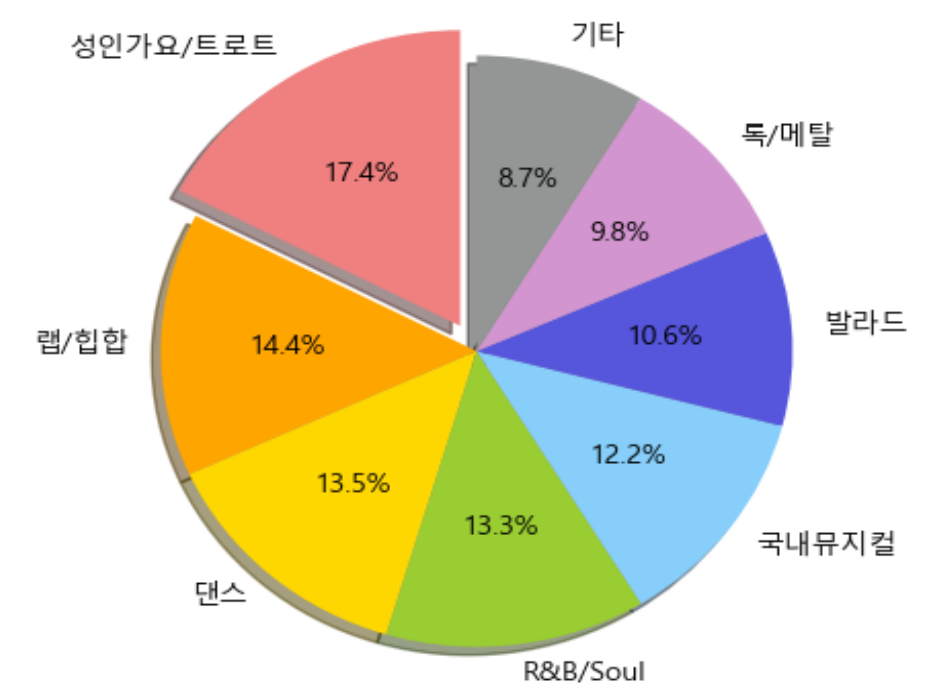
2000년 장르분포



2010년 장르분포



2020년 장르분포





## 국내 뮤지컬









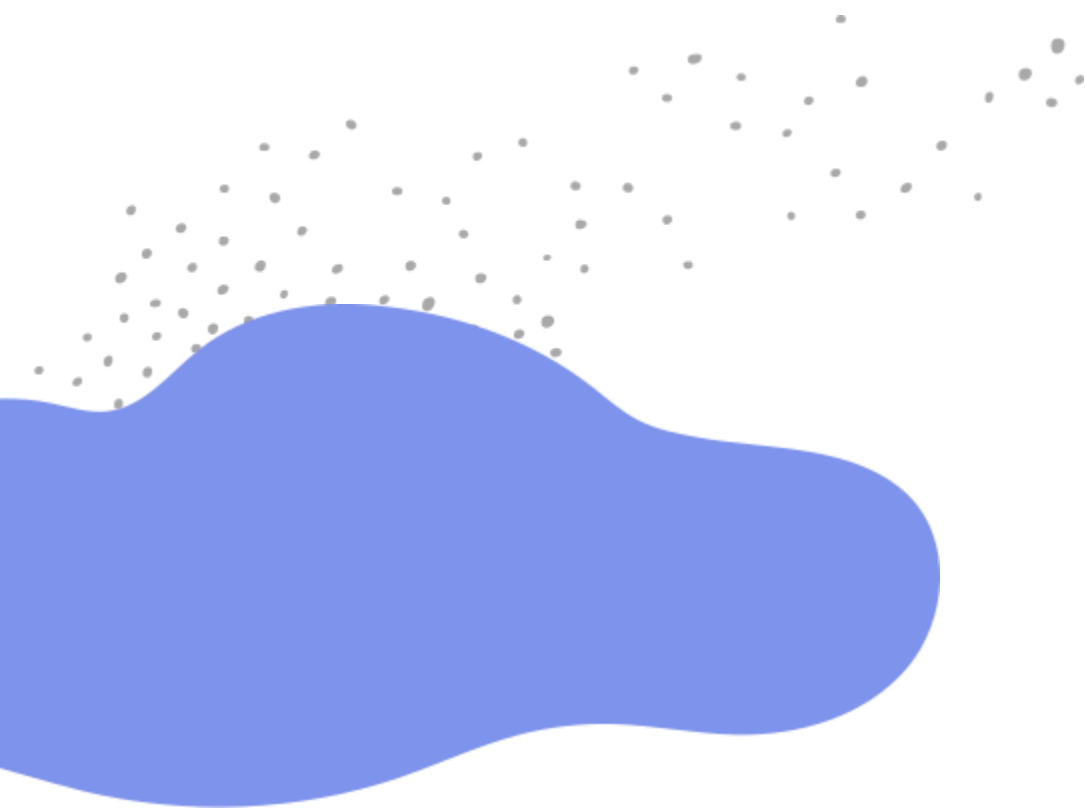
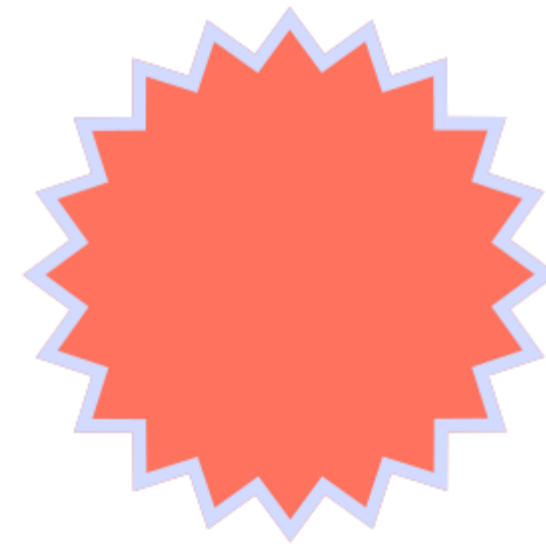
## 플레이 리스트에 가장 많이 들어있는 곡

순위		곡명	가수	앨범	발매일	플레이리스트에 들어간 수
1		아로하	조정석	슬기로운 의사생활 OST Part 3	2020.03.27	107
2		Ditto	NewJeans	NewJeans 'OMG'	2022.12.19	93
3		Hype boy	NewJeans	NewJeans 1st EP 'New Jeans'	2022.08.01	91
4		I AM	IVE (아이브)	I've IVE	2023.04.10	88
5		밤편지	아이유	밤편지	2017.03.24	88
6		After LIKE	IVE (아이브)	After LIKE	2022.08.22	85
7		첫눈처럼 너에게 가겠다	에일리(Ailee)	도깨비 OST Part.9	2017.01.07	84
8		Candy	NCT DREAM	Candy - Winter Special Mini Album	2022.12.16	83
9		사랑인가 봐	멜로망스	사랑인가 봐 (사내맞선 OST 스페셜 트랙)	2022.02.18	81
10		OMG	NewJeans	NewJeans 'OMG'	2023.01.02	80



# 3. 데이터 처리 및 개발 과정

- 추천 시스템
- 웹 개인화



# 컨텐츠 기반

## ★ 가사 학습전 불용어 처리

```
# 한글 불용어 처리
with open('data/불용어.txt') as st:
    lines = st.readlines()
stop_words = [line.split('\t')[0] for line in lines]
stop_words.extend('은 는 를 도 을 며 의 에 게 니 거 로 요 과 래 랑 파 여 에게'.split())
```

```
from konlpy.tag import Okt
okt = Okt()
```

```
lyrics = []
for lyric in df.lyric:
    lyric = lyric.replace('\n', ' ')
    morphs = okt.morphs(lyric, stem=True)
    tmp = [word for word in morphs if word not in stop_words]
    lyrics.append(' '.join(tmp))
df['morphs'] = lyrics
```

# 컨텐츠 기반

## ★ TfidfVectorizer 사용

- 유사도 높은 노래가 나올 수 있도록

전처리한 가사(morphs), 제목, 아티스트 2, 작곡가 2, 작사가 2, 장르 3 => 학습

```
df['total'] = df.morphs + (' ' + df.title) + (' ' + df.artist) * 2 + (' ' + df.composer) * 2 + (' ' + df.lyricist) * 2 + (' ' + df.genre) * 3
```

```
from sklearn.feature_extraction.text import TfidfVectorizer
tvect = TfidfVectorizer(stop_words='english')
total_tv = tvect.fit_transform(df.total)
from sklearn.metrics.pairwise import cosine_similarity
cosine_sim = cosine_similarity(total_tv)
```

★ Sim 모듈 생성한 후 파일로 저장  
웹 구현 시 빠른 결과 화면 출력

## ★ 코사인 유사도 측정

자기 자신을 제외한 후 유사도 순으로 TOP 5 출력

```
from sklearn.metrics.pairwise import linear_kernel
cosine_sim_total = linear_kernel(total_tv, total_tv)
sim_scores = pd.Series(cosine_sim_total[indices['32720013']])
sim_scores.sort_values(ascending=False).head(6).tail(5)
```

count	6243.000000
mean	0.345133
std	0.163045
min	0.039842
25%	0.231017
50%	0.302136
75%	0.430867
max	0.971573

# 숨은명곡 (좋아요&댓글 기반)

## ★ 콘텐츠 기반 추천 시스템에 4분위 접목

- 좋아요와 댓글수를 합산한 기록치가 하위 25%~50%에 해당한다면 유명하진 않아도 매니아층이 있는 노래로 판별

```
# 찾고 싶은 구간 정하기
df['comment_like_total'] = df.comment + df.like
numbers = df['comment_like_total']
sorted_numbers = np.sort(numbers)
q1 = np.percentile(sorted_numbers, 25)
q2 = np.percentile(sorted_numbers, 50)
filtered_data = df[(df['comment_like_total'] >= q1) & (df['comment_like_total'] < q2)]
filtered_data = filtered_data[['songId', 'comment_like_total']]
filtered_data.songId.values
```

# 플레이리스트 기반

★ 검색한 곡이 속해있는 플레이리스트를 찾아 비슷한 노래를 추천 (자신 제외)

플레이리스트에 함께 들어있던 노래들이에요!



불장난  
BLACKPINK



달라달라  
ITZY (있지)



IDOL  
방탄소년단



TT  
TWICE (트와이스)



Lion Heart  
소녀시대 (GIRLS' GENERATION)



# 현재 날씨 기반



Group 800: Clear

ID	Main	Description	Icon
800	Clear	clear sky	● 01d ● 01n

Group 80x: Clouds

ID	Main	Description	Icon
801	Clouds	few clouds: 11-25%	☀ 02d ☾ 02n
802	Clouds	scattered clouds: 25-50%	☁ 03d ☁ 03n
803	Clouds	broken clouds: 51-84%	☁ 04d ☁ 04n
804	Clouds	overcast clouds: 85-100%	☁ 04d ☁ 04n








# 현재 날씨 기반

★ 사용자 위치 날씨 & 플레이리스트에 들어있는 태그와 결합하여 플레이리스트 산출 (곡 랜덤)

현재 날씨는 "온흐림 ☁️" 입니다.

## # 지금 듣기 딱 좋은 # Playlist



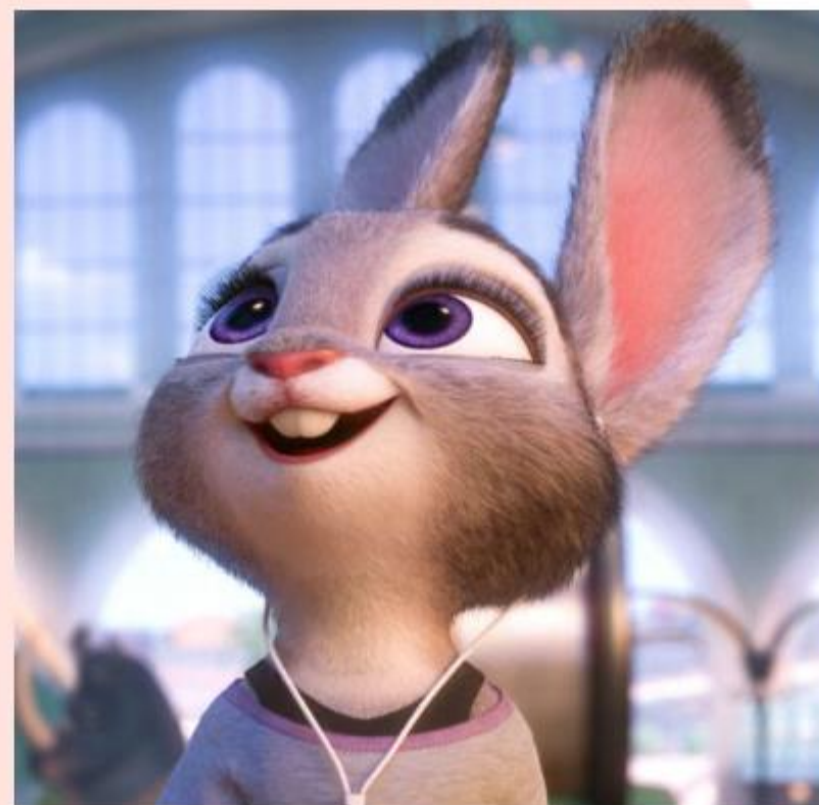
	존재만으로	원슈타인	스물다섯 스물하나 OST Part 4	♥
	침식	Room306	술과 꽃	♥
	Shut Up And Let Me Go	The Ting Tings	Shut Up And Let Me Go (Maxi Single)	♥
	281.31km ( To. )	김뮤지엄 (KIMMUSEUM)	281.31km EP	♥
	시간을 돌려서	소유 (SOYOU)	조선변호사 오리지널 사운드 트랙 4	♥

# 현재 시간 기반

★ 사용자 접속 시간 & 플레이리스트에 들어있는 태그와 결합하여 플레이리스트 산출 (곡 랜덤)

현재 시간은 10시 35분 입니다.

## # 지금 듣기 딱 좋은 # Playlist



	으르렁 (Growl)	EXO	The 1st Album 'XOXO' Repackage	♥
	Movie Star	미주 (MIJOO)	Movie Star	♥
	주라주라	둘째이모 김다비	주라주라	♥
	Feel Special	TWICE (트와이스)	Feel Special	♥
	고민보다 Go	방탄소년단	LOVE YOURSELF 承 'Her'	♥

# 웹 개인화 과정

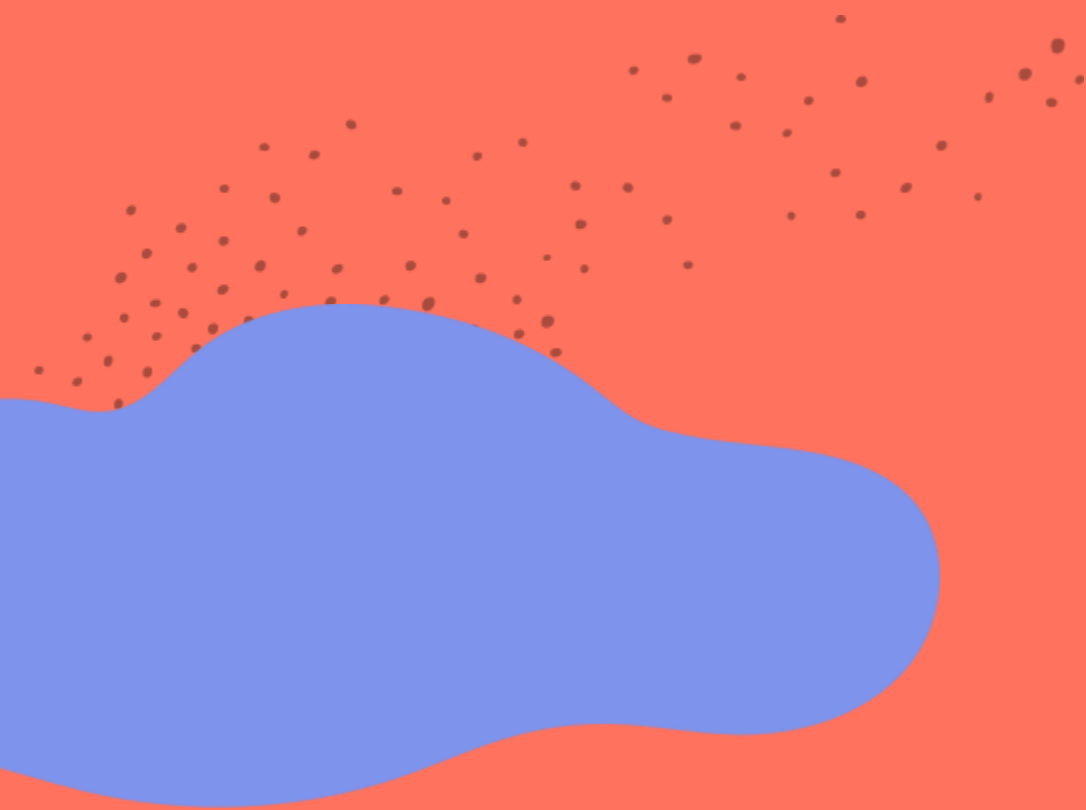
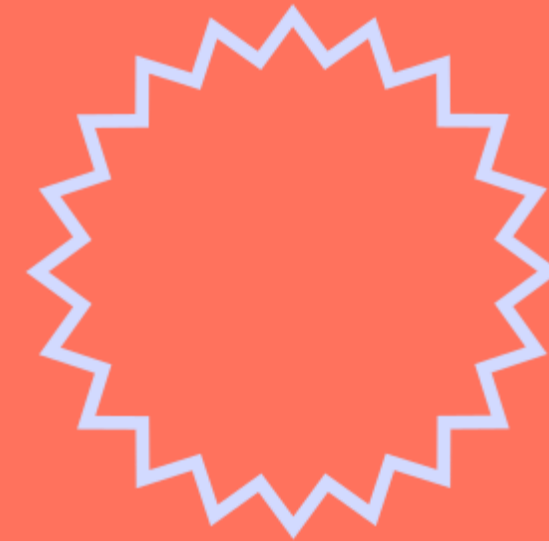
## ★ Sqlite3 사용하여 DB 생성 후 Server 연결

사용자				
user				
uid	사용자ID	text	primary key	not null
uname	이름	text		not null
pwd	패스워드	text		not null
birthday	생년월일	int		not null
gender	성별	text	1:남자, 2:여자	
email	이메일	text		not null
regdate	가입날짜	Date	기본값: CURRENT_DATE	not null
사용자 이력				
user_searched				
sid	순서	integer	primary key/auto	not null
uid	사용자ID	text	foreign key	not null
stime	검색시간	text	ex)2023.06.05 13:34:09	not null
songid	검색곡	text		
img	곡이미지	text		
title	제목	text		
artist	아티스트	text		
album	앨범	text		

	uid	uname	pwd	birthday	gender	email	regdate
	필터	필터	필터	필터	필터	필터	필터
1	maria	마리아	A6xnQhzbz4Vx2HuGI4IXwZ5U2i8iziLRFnhP5eNf...	2023-06-06	2	dfdfd@dfdf	2023-06-05
2	yumii2307	김유미	jn1G642GfcZE7TUPJW7YLE/F6GeOGbHvZR3/...	2000-01-20	2	yumii2307@naver.com	2023-06-08
3	smpet	김화영	pmWkWSBCL51Bfkhn79xPuKBKHz//...	1976-12-21	2	ssmpet@naver.com	2023-06-08
4	gaga	최가람	pmWkWSBCL51Bfkhn79xPuKBKHz//...	1994-06-24	2	dfdfd@dfdf	2023-06-08
5	james	제임스	A6xnQhzbz4Vx2HuGI4IXwZ5U2i8iziLRFnhP5eNf...	2023-05-08	1	jamesxxx@gamil.com	2023-06-08



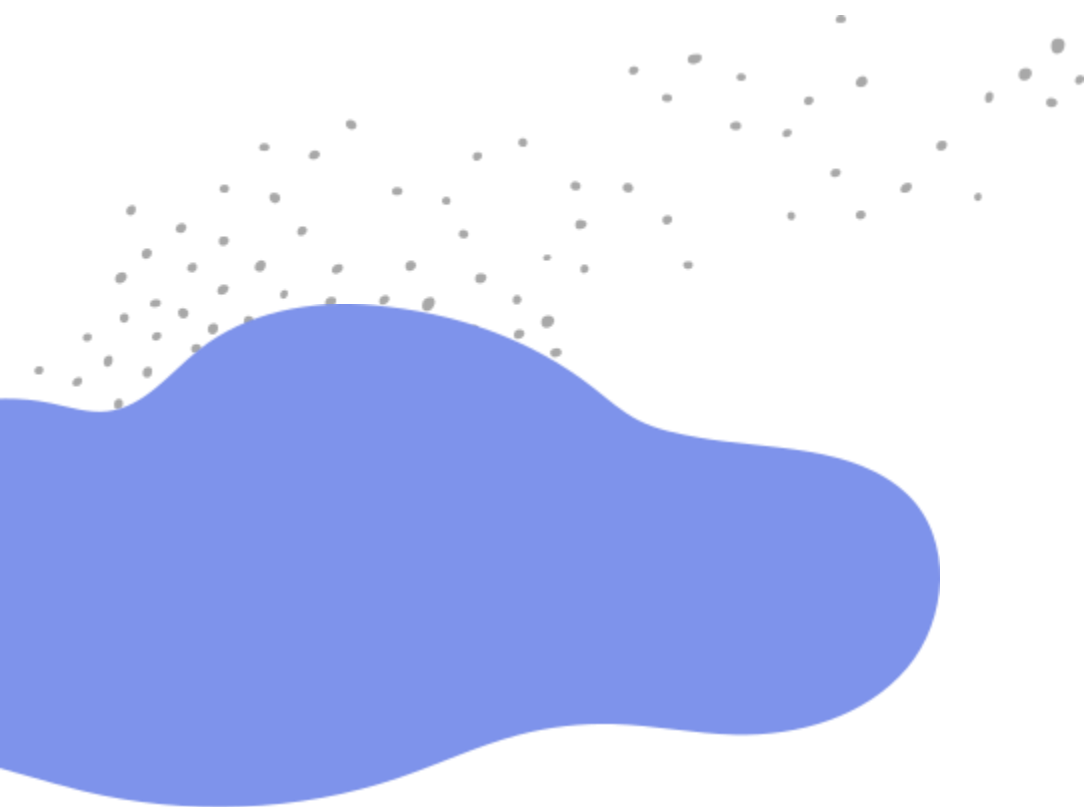
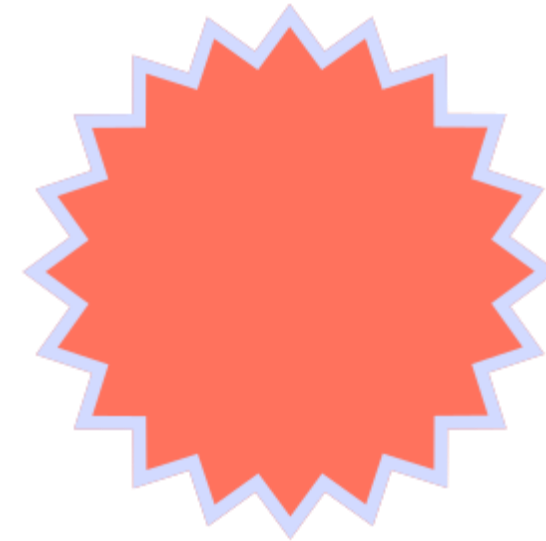
# WEB TIME





## 4. 프로젝트 리뷰

- 기대효과
- 가능성 및 한계
- 소감





# 기대효과

그림 2-2-2-1 음악 스트리밍 서비스 이용 빈도

(Base: 음악 스트리밍 서비스 이용자, 단위: %)

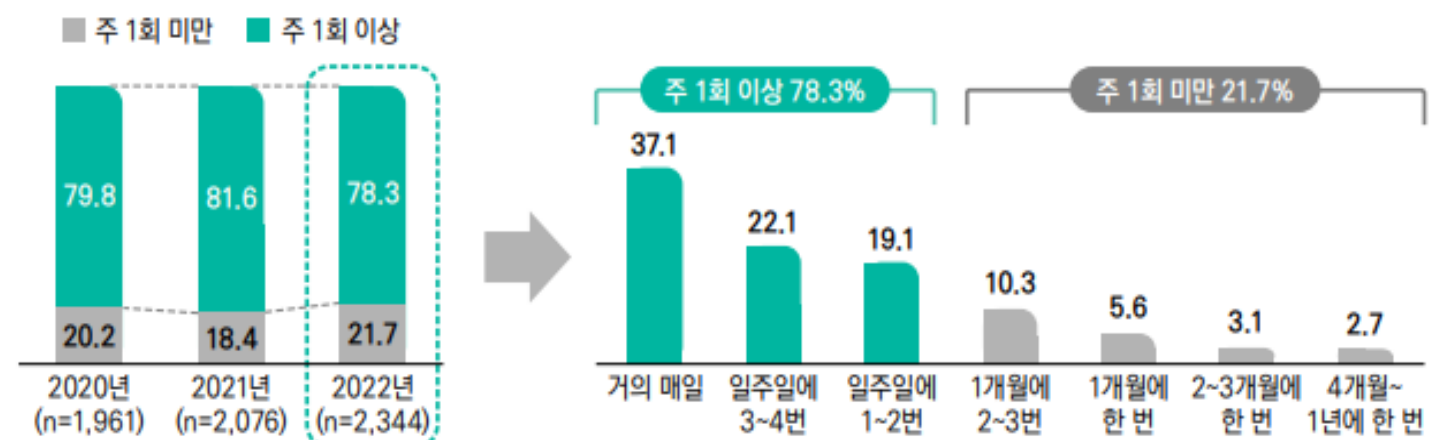


표 2-2-2-2 응답자 특성별 음악 스트리밍 서비스 이용 방법

(Base: 음악 스트리밍 서비스 이용자, 중복 응답, 단위: %(1+2순위 기준))

구분	사례 수	그때그때 듣고 싶은 곡이나 앨범을 직접 검색해서 감상	내가 편집/저장해놓은 음악 감상	음악 서비스가 제공하는 차트 (예. Top 100) 에서 듣고 싶은 곡이나 앨범을 선택해서 감상	서비스가 제안하는 '테마리스트', '내 취향 맞춤형' 등 선곡리스트 감상	스타 DJ/일반 이용자 DJ 등의 플레이리스트 감상	기타
전체	(2,344)	65.3	46.8	44.0	21.2	5.8	0.3
성별	남성 (1,182)	65.1	44.5	44.8	22.3	5.8	0.2
	여성 (1,162)	65.6	49.2	43.1	20.2	5.8	0.3
연령별	10대 (312)	64.7	44.2	46.2	22.4	4.8	0.6
	20대 (506)	59.9	56.1	40.9	20.6	5.1	0.0
	30대 (515)	62.9	49.9	45.8	18.4	4.7	0.6
	40대 (426)	65.3	41.8	49.5	21.8	8.7	0.0
	50대 (390)	71.3	41.8	42.3	22.6	3.8	0.3
	60대 (195)	74.9	40.0	34.9	24.6	9.7	0.0

다음 차트 출처 : 한국콘텐츠진흥원 2022년 음악산업백서

## ★ 음악 스트리밍 서비스 이용 빈도

- 주 1회 이상 이용 (78.3%)

## ★ 이용 방법

- 전 연령층이 그때그때 듣고 싶은 곡이나 앨범을 직접 검색 (65.3%)

### 1) 사용자 취향에 맞는 노래 검색

- 찾는 시간 단축

### 2) 사용자에게 새로운 음악 추천

- 음악 감상의 즐거움 제공

### 3) 새로운 아티스트 / 작곡가

- 음악에 대한 노래 홍보 기회 제공

# 가능성 & 한계

## 가능성

검색기록 최근 9개까지 출력 => 더보기 기능을 추가하여 전체 기록을 조회할 수 있다. (추후 협업 필터링 가능)

개인정보수집 동의를 얻어 후에 사용자 정보 통계까지 진행할 수 있다.

## 한계

데이터셋의 분포가 다채롭지 못하고 지속적인 업데이트의 한계가 있다.

서버 생성 및 서비스배포의 저작권 문제가 있다.

다국적 노래 서비스가 가능하나 데이터 처리 과정에 문제 발생 요지가 있다. (다국적 직원 필요)

# 소감



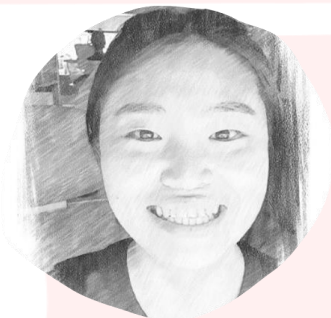
김유미

기획과 보여지는 웹에만 치중하고 있었는데 서버 연결을 통해 생각했던 이미지가 구현이 되는게 신기했습니다. 그래서 서버에 관련된 백엔드 기술에 매력을 많이 느끼는 계기가 되었습니다.



김화영

데이터가 만만치 않았지만 이상치가 나오거나 오류가 생겨도 조원들과 같이 수정하면서 해 가는 과정이 재미있었고 빅데이터에 대한 관심도가 예전보다 높아졌으며, 가장 기본인 데이터가 중요하다는 것을 더욱 알게 되었습니다.



최가람

데이터셋 처리, 코드 작업, 오류 예측 과정에 꼼꼼함이 무척이나 중요하다고 느꼈습니다. 그만큼 팀워크도 중요한 요소임을 느껴 열심히 해서 도움이 되면 좋겠다고 생각했습니다.

# THANK YOU

발표를 들어주셔서 감사합니다 :)

