데이터 분석 기획서

I 기획서 개요

○ 분석 주제명

- 최근 10년 트렌드분석을 통한 'BTS'의 '빌보드TOP20' 전략적 접근

○ 역할 및 소요시간

< 역할 >
- 한찬희(PM) : 기획
- 김예지 : 데이터 모델링
- 김성수 : 데이터 전처리
- 황두경 : EDA
- 김랑하 : 데이터수집, 분석

< 소요시간 >

단계	세부정보	10/16(수)	10/17(목)	10/18(금)	10/21(월)	10/22(화)	10/23(수)	10/24(목)	10/25(금)	10/28(월)	10/29(화)
프로젝트 기획	기획서 작성										
데이터 수집	빌보드 차트 데이터 수집										
	음악 정보 데이터 수집										
데이터 가공	데이터 전처리										
	데이터 1차 분석										
데이터 분석	EDA										
	데이터 분석										
모델링	데이터 시각화										
	베이스라인 설정										
프로젝트 마무리	실험 및 모델 해석										

○ 목차

○ 배경 및 필요성

- 전 세계 음악 시장은 매년 지속적인 성장을 이루고 있으며, **2023년에는 286억 달러(한화 약 39.6조 원)의 규모에 도달**했습니다. 특히 2**022년에서 2023년 사이에는 "10.2%"의 높은 성장률**을 기록하며, 음악 시장은 거대 산업으로 평가되고 있습니다. 이 성장은 주로 **유료 스트리밍 서비스 가입자의 증가로 높은 성장률에 매우 큰 기여도를 주었습니다.**
이를 통해, 글로벌 음악시장내에서도 **가장 두드러지게 비중이 높은 스트리밍 시장에 데이터분석 기반 새로운 시스템을 제공**하며, 전통적인 A&R(Artists and Repertoire) 방식에서 벗어나, 변화하는 시장 환경에 적응한 새로운 비즈니스 전략이 요구되고 있습니다. 스트리밍, 소셜 미디어, 그리고 **데이터 기반의 음악 소비 트렌드는 더 이상 기존의 방식으로 아티스트를 발굴하고 매니지먼트 하는 방식에서 벗어나 새로운 시스템과 비즈니스, 마케팅 전략을 제시할 것이라 예상합니다.**

이에 따라, 빌보드 HOT100과 같은 세계에서 가장 영향력 있는 음원 차트의 트렌드를 분석함으로써, 빌보드 HOT100 차트에 진입하거나 TOP10에 올라갈 수 있는 곡에 대한 인사이트를 얻어 더욱 구체적이고 효과적인 마케팅 전략을 설계할 수 있을 것입니다.

II 데이터 수집

○ 데이터탐색

- 빌보드 HOT100사이트에서 2014년부터 2023년의 노래제목, 가수이름을 WEEKLY단위로 1~100위
- 변수설정
: Year(연도), Month(월), Week(주차), Rank(순위), Title(노래제목), Artist(가수이름), Lyrics(노래가사), Genre(장르), BPM, Duration_sec(노래길이), color1, color2, color3, rgb1, rgb2, rgb3, Featuring(피쳐링여부(1,0))

○ 데이터 수집 방법

- Title, Artist, Year, Month, Week, Rank, Featuring : '빌보드HOT100' 웹페이지에서 2014년 1월 1주차 ~ 2023년 12월 마지막 주까지 크롤링
- BPM, Duration_sec : spotify API를 이용해서 크롤링
- Genre : 멜론뮤직 웹사이트에서 검색창에 'Title-Artist' 형식으로 검색 후 첫 번째 결과 곡정보 클릭 후 장르부분 크롤링
- Lyrics : OVH API, Genie음원사이트 크롤링, Google검색 크롤링
- 대표색상 및 RGB값(color1, color2, color3, rgb1, rgb2, rgb3) : color thief라이브러리 사용

Ⅲ 데이터 분석

○ 탐색적 데이터 분석(EDA)

<사용Tool 및 분석방법>
- 데이터의 특징을 파악하는 시각화 진행
- 시각화 도구 : Python 기반의 matplotlib, Color Thief, WordCloud 라이브러리
- 분석 기법 : TF-IDF, RAKE 알고리즘, KeyBERT 등을 사용

<시각화 및 추가분석>
- 연도별로 데이터의 특징을 파악하기 위해 장르, 노래 길이, BPM, 피처링 여부, 가사 등의 변수를 분기, 6개월, 1년 단위로 시각화
- 단순 데이터의 특징을 파악하기 어려운 가사 분석에는 WordCloud와 VADER 라이브러리를 사용하여 단순 빈도 및 PCA를 통한 시각화 및 가사 감정 분석 수행
- 앨범 커버 대표 색상 분석에는 Color Thief 라이브러리와 HSL 기반 세부 범주화를 적용하여 시각화

Ⅳ 모델링

○ 알고리즘

- 트리 기반 : 랜덤 포레스트(Random Forest), 디시전 트리(Decision Tree)
- 경사하강법 기반 : 로지스틱 회귀분석(Logistic Regression)
- 확률 기반 : 나이브 베이즈 분류(Naive Bayes Classifier)
- 거리 기반 : SVM(Support Vector Machine), KNN(K-Nearest Neighbors)
- 수치형 데이터에 대해서만 다중공선성을 확인.

Ⅴ 기대효과 및 발전방향

- 분석된 데이터를 기반으로 빌보드 차트 진입 가능성을 예측하고, 효과적인 마케팅 전략과 접근 방안을 제시할 수 있습니다. 이를 통해 BTS와 같은 아티스트들이 글로벌 차트에서 성공할 수 있는 구체적이고 실용적인 전략을 구축하는 데 기여할 것으로 기대됩니다.