

# 음악 추천 알고리즘에 탑승하기

- Spotify music data analysis -

4조 박장호

# CONTENTS

---

1

Intro

2

Dataset

Data source  
Data features  
Data analysis  
environment

3

EDA

Feature  
Correlation  
Genre feature  
Time series  
Artist  
Popularity

4

Outro

Conclusion  
Feedback  
Reference



# 1. Intro

## 1. Intro

---

동영상 스트리밍 ->



음악 스트리밍 ->



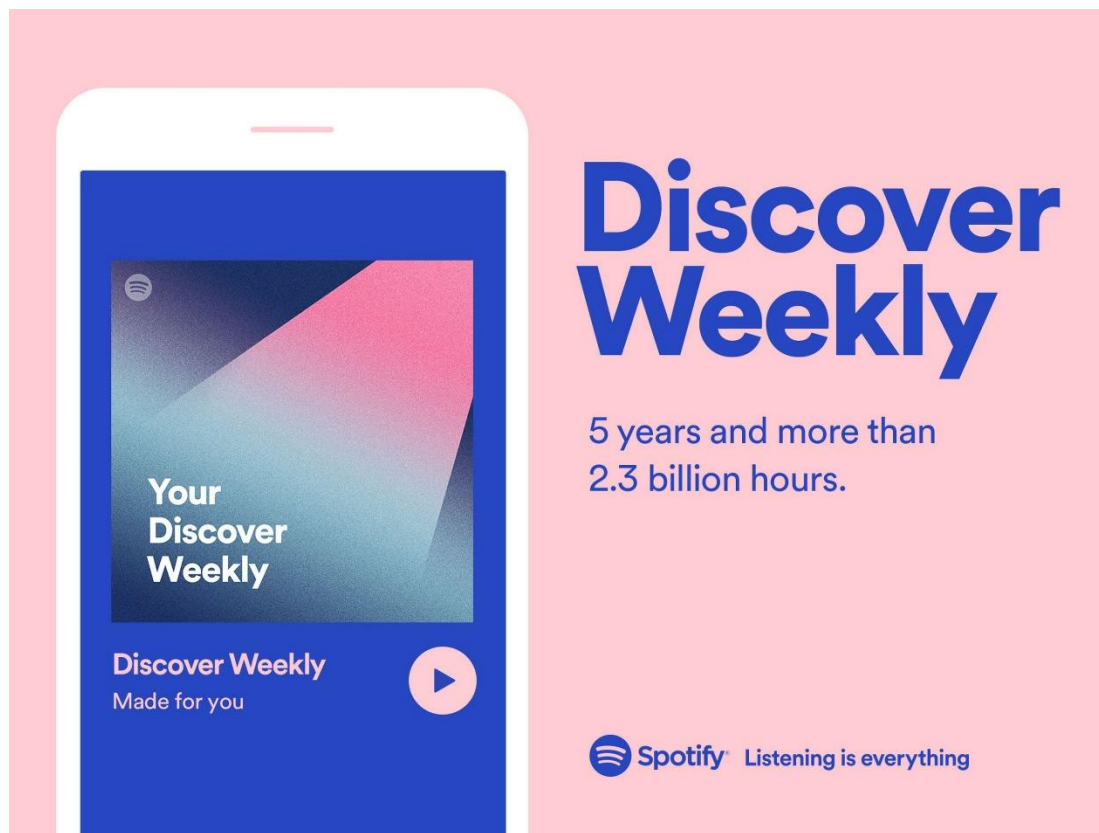
# 1. Intro

제 꿈은...



작곡!

# 1. Intro



독보적인 음악 추천 알고리즘

# 1. Intro



OTM Conference  
OTM 2019

An  
Ann

Authors

J. García

Conferen  
First On

# Spotify의 data는 어떤 모습일까?

# 어떤 connection을 만들 수 있을까?



Search &  
Recommendation

Learning



User Modeling

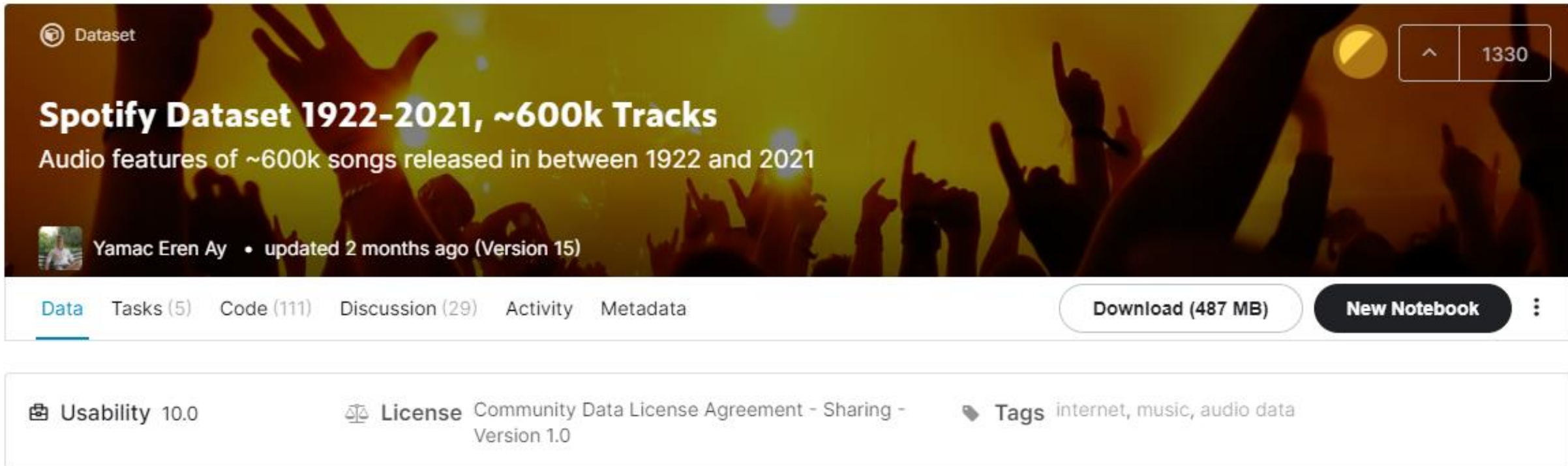


## 2. **Dataset**



## 2. Dataset

Data source



The screenshot shows the Kaggle dataset page for "Spotify Dataset 1922-2021, ~600k Tracks". The header features a banner with silhouettes of hands raised in the air. Below the banner, the dataset title and description are displayed. The creator's name and update status are shown. A navigation bar includes links for Data, Tasks, Code, Discussion, Activity, and Metadata. Action buttons for downloading and creating a notebook are present. A metadata section at the bottom provides details on usability, license, and tags.

**Spotify Dataset 1922-2021, ~600k Tracks**  
Audio features of ~600k songs released in between 1922 and 2021

Yamac Eren Ay • updated 2 months ago (Version 15)

[Data](#) [Tasks \(5\)](#) [Code \(111\)](#) [Discussion \(29\)](#) [Activity](#) [Metadata](#) [Download \(487 MB\)](#) [New Notebook](#)

**Usability** 10.0 **License** Community Data License Agreement - Sharing - Version 1.0 **Tags** internet, music, audio data

**Spotify Dataset 1922-2021, ~600k Tracks (Kaggle)**  
**-Audio features of ~600k songs released in between 1922 and 2021**

**Spotify Web API, Spotipy (Python module for Spotify Web Server)**

## 2. Dataset

Data source

### Data Explorer

487.48 MB

- artists.csv
- data\_by\_artist\_o.csv
- data\_by\_genres\_o.csv
- data\_by\_year\_o.csv
- data\_o.csv
- {i} dict\_artists.json
- tracks.csv



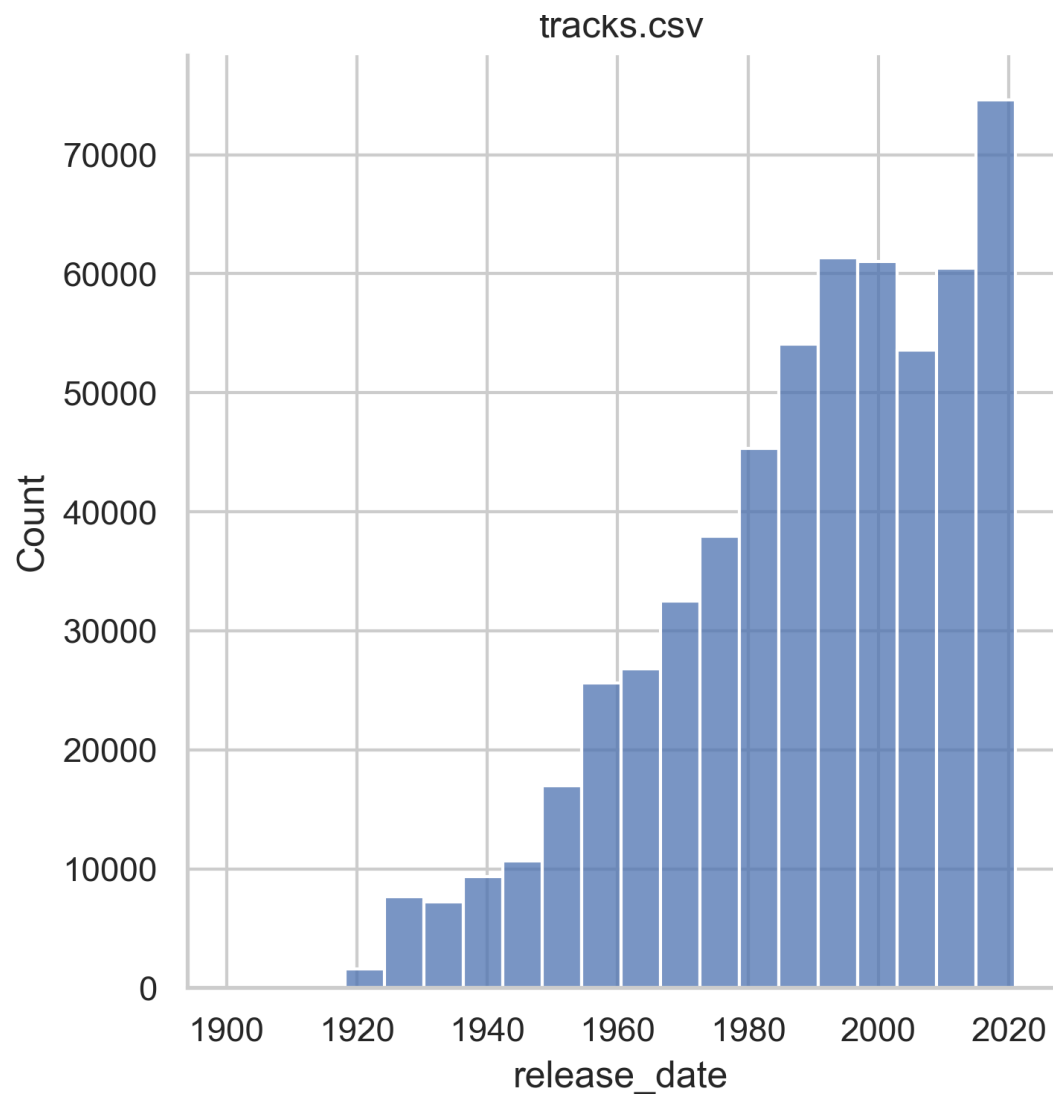
## 2. Dataset

## Data feature

[illegible]

## 2. Dataset

### Data feature



### track.csv

- Rows : 586,672
- Columns : 20
- Due : 1922~2021.04

## 2. Dataset

### Data feature

- acousticness : 어쿠스틱 정도 (0.0~1.0)
- danceability : 댄스에 적합한 정도 (0.0~1.0)
- energy : 격렬하고 활동적인 정도. 빠르고, 소리가 큰 경향 (0.0 ~ 1.0)
- instrumentalness : 보컬 유무 (1.0에 가까울수록 instrumental) (0.0 ~ 1.0)
- liveness : 음원에 관객 소리가 있는 정도. 0.8 이상 시 라이브 음원으로 판단 가능 (0.0 ~ 1.0)
- speechness : 목소리 정도를 감지. 토크쇼 or 오디오북은 1.0, 0.66 이상은 대부분 구어, 0.33~0.66은 음악과 구어(랩 포함), 0.33 미만은 대부분 음악이나 비 언어적 트랙 (0.0 ~ 1.0)
- valence : 긍정적인 정도 (0.0 ~ 1.0)
- tempo : 평균 beats per minute (bpm)
- duration\_sec : 플레이 타임 (초)
- loudness : 트랙 전체 소리(dB) 평균화된 트랙 음량을 상대적으로 비교. (-60dB~0dB)
- mode : major(1) 혹은 minor(0)
- key : 0 - C , 1 - C#, 2 - D ... (0~11)
- Popularity : 유명한 정도. 트랙이 플레이 된 횟수 (0~100)

## 2. Dataset

### Data analysis Environment

---

- Jupyter Notebook (Visual Studio Code)
- Python 3.7.3
- Pandas
- Numpy
- Matplotlib
- Plotly
- Seaborn



# 3. EDA

### 3. EDA

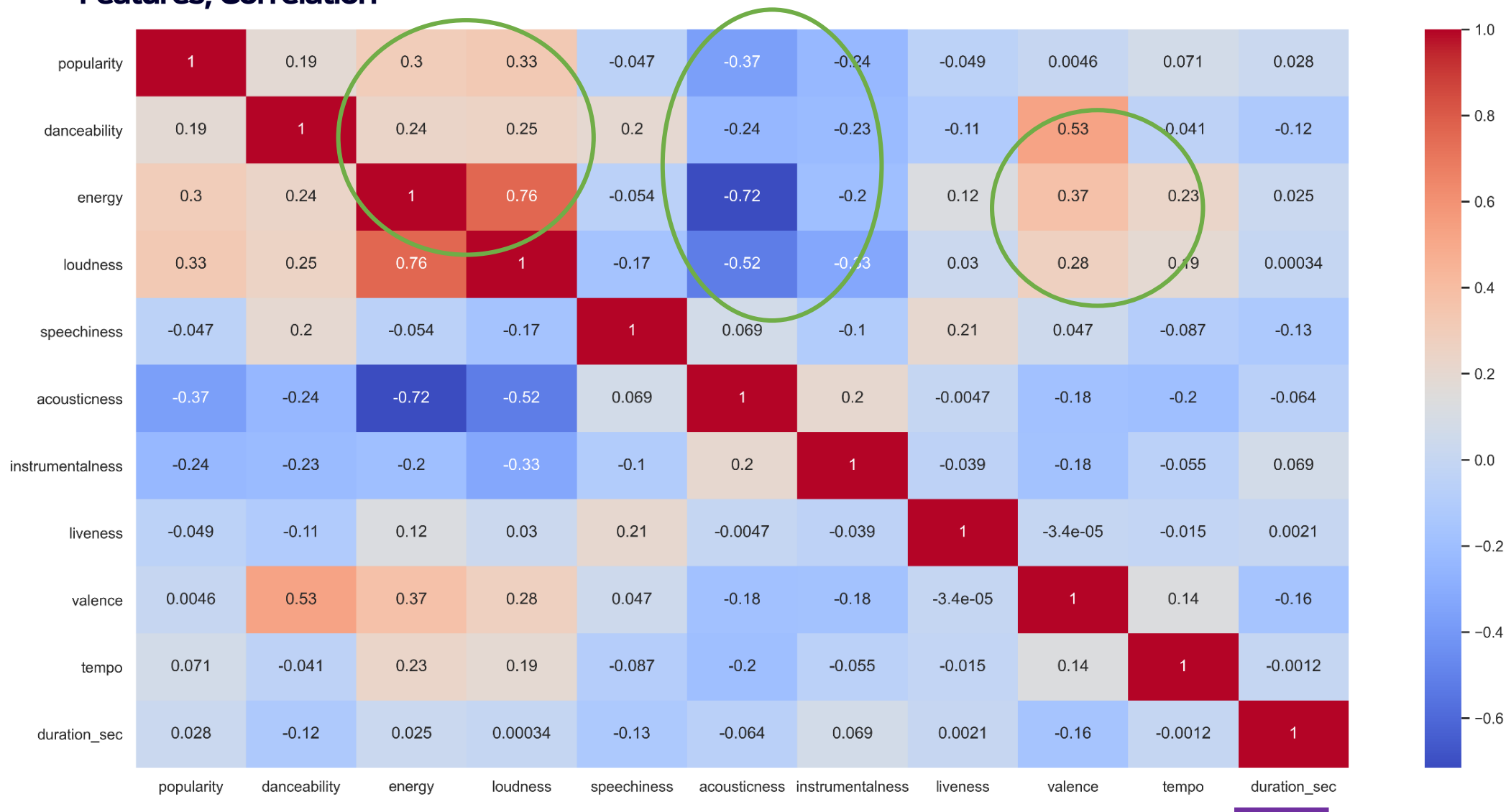
---

1. 음악의 특징들, 상관관계
2. 장르별 음악의 특징
3. 시간대에 따른 음악의 변천사
4. 가수들의 Popularity



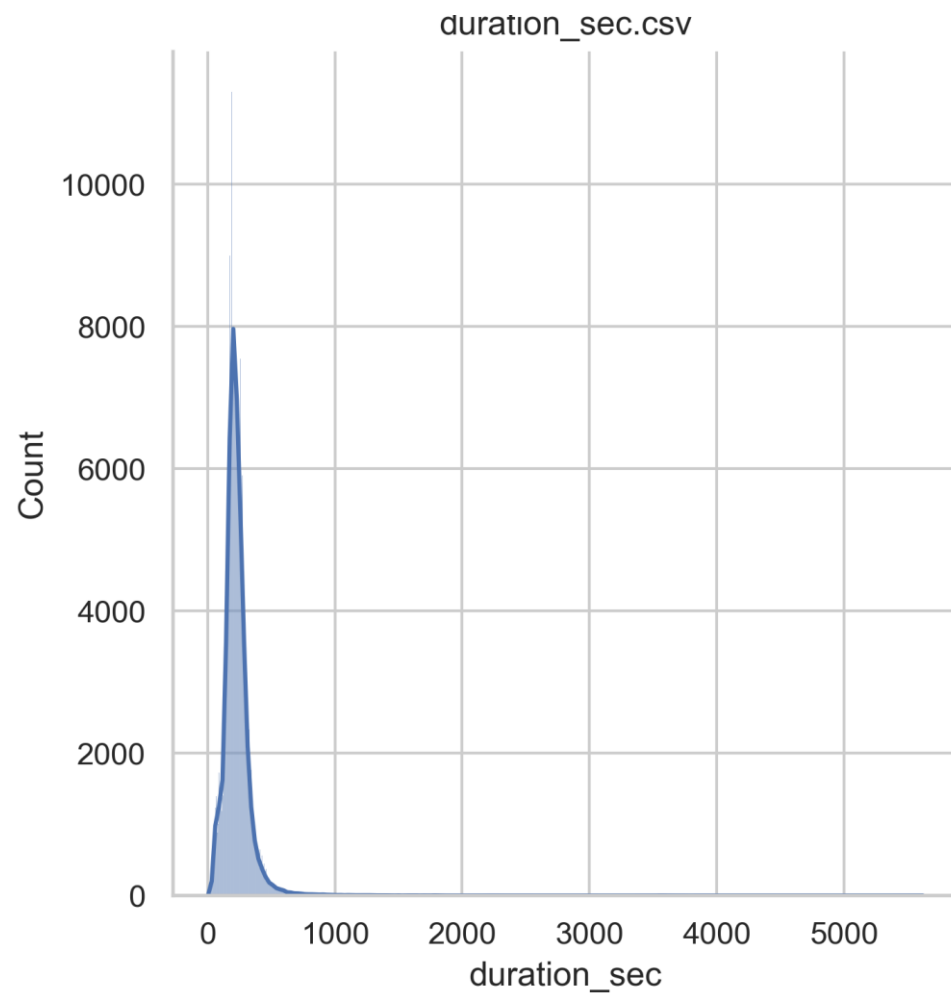
### 3. EDA

#### Features, Correlation



### 3. EDA

#### Features, Correlation

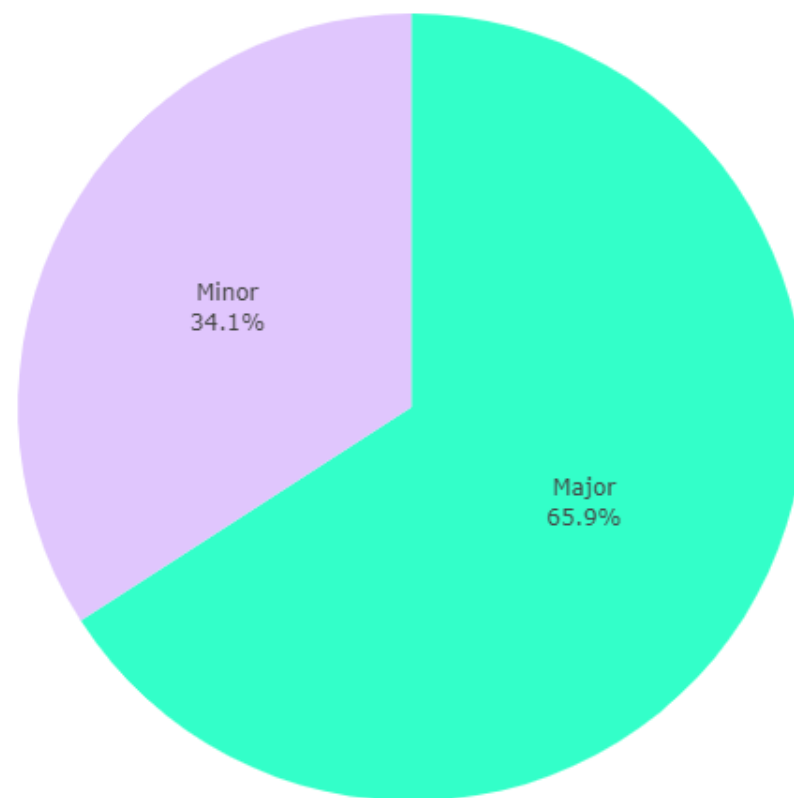
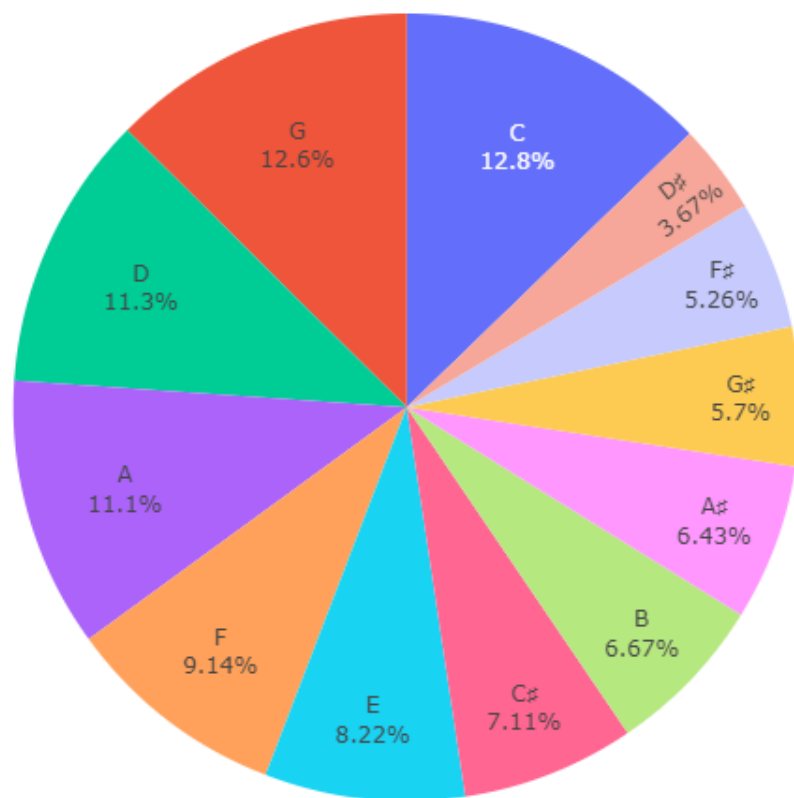


Mean : 230 sec

### 3. EDA

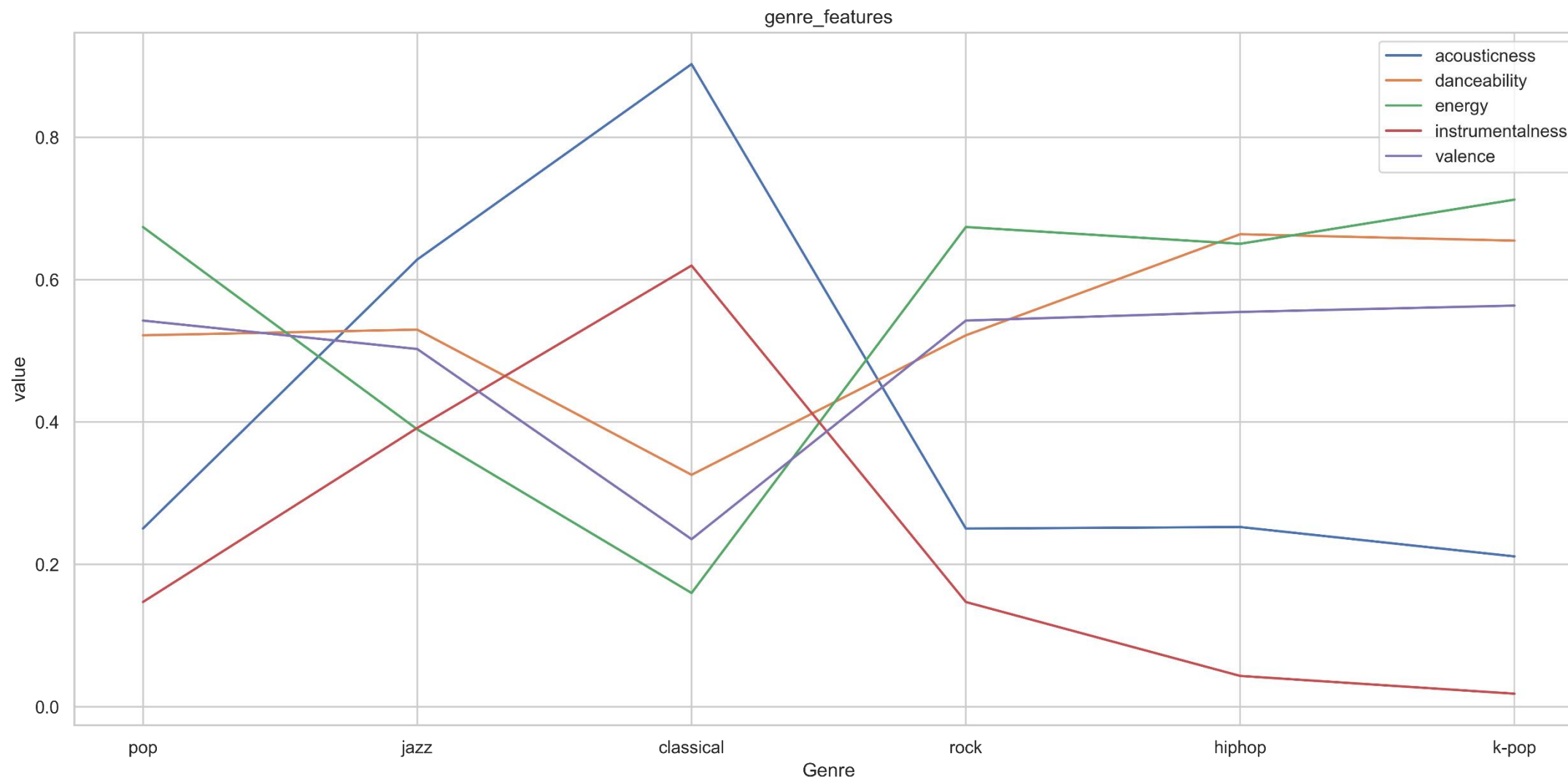
#### Features, Correlation

Keys / Modes



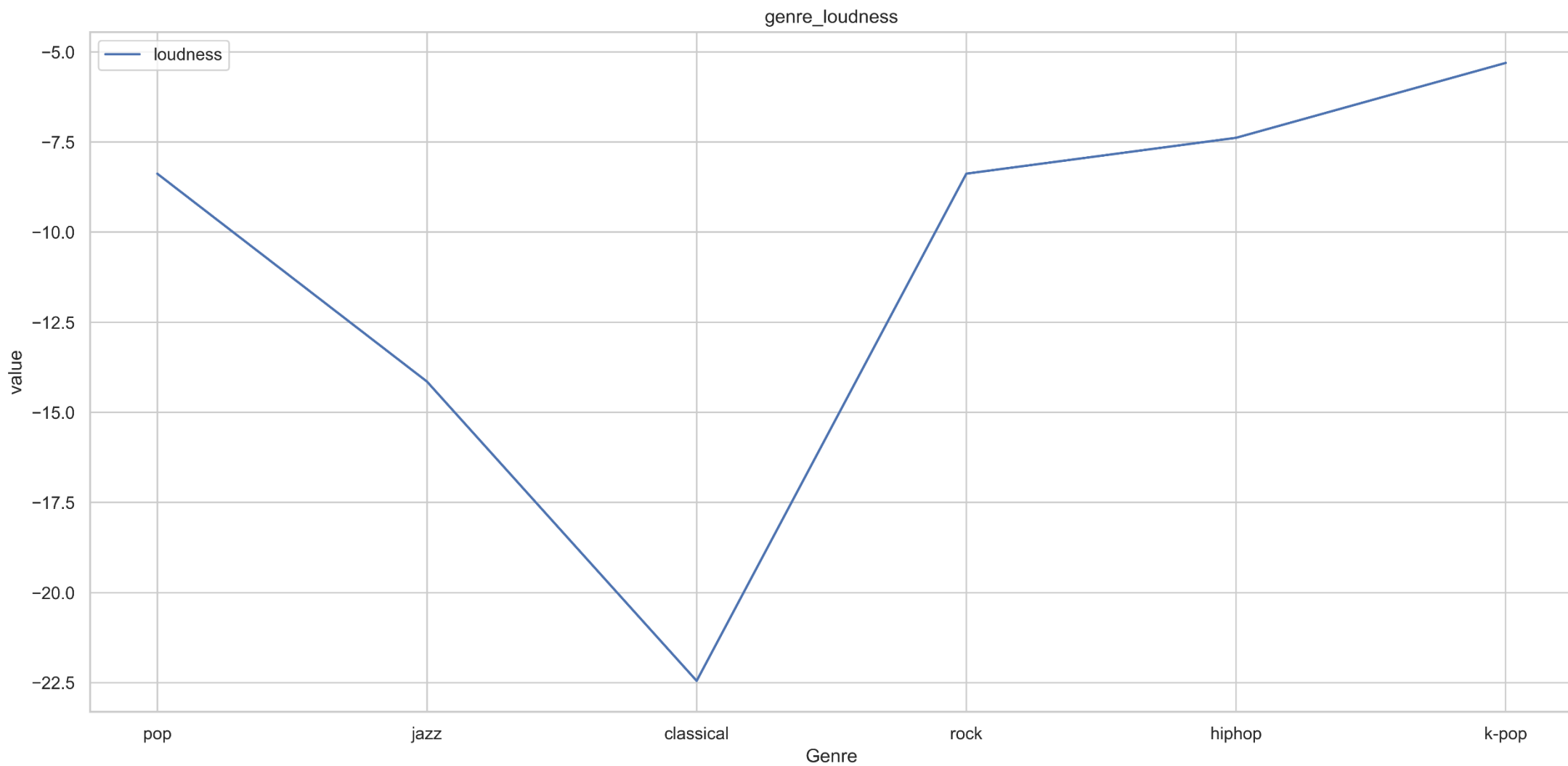
### 3. EDA

#### Genre Feature



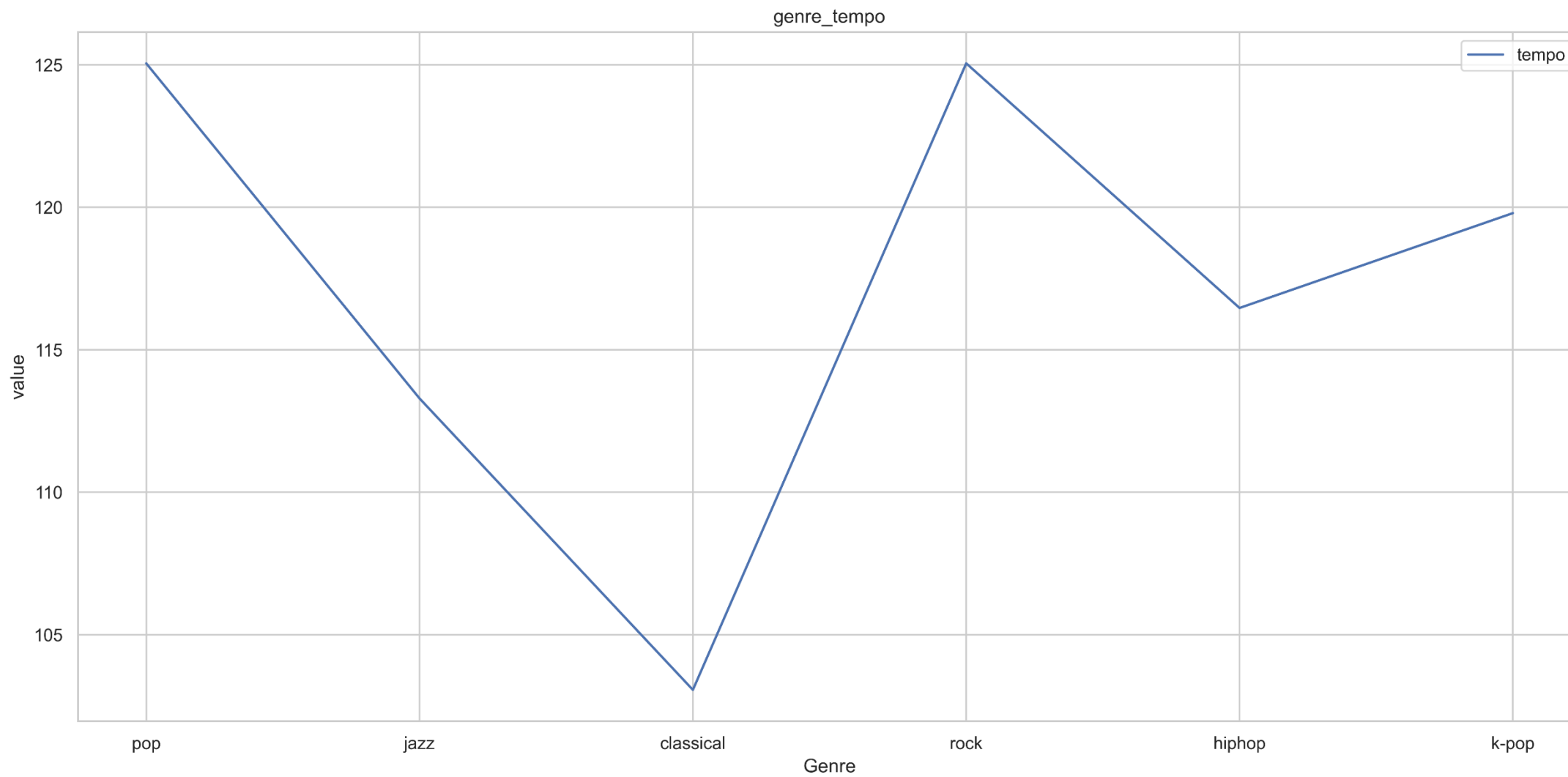
### 3. EDA

#### Genre Feature



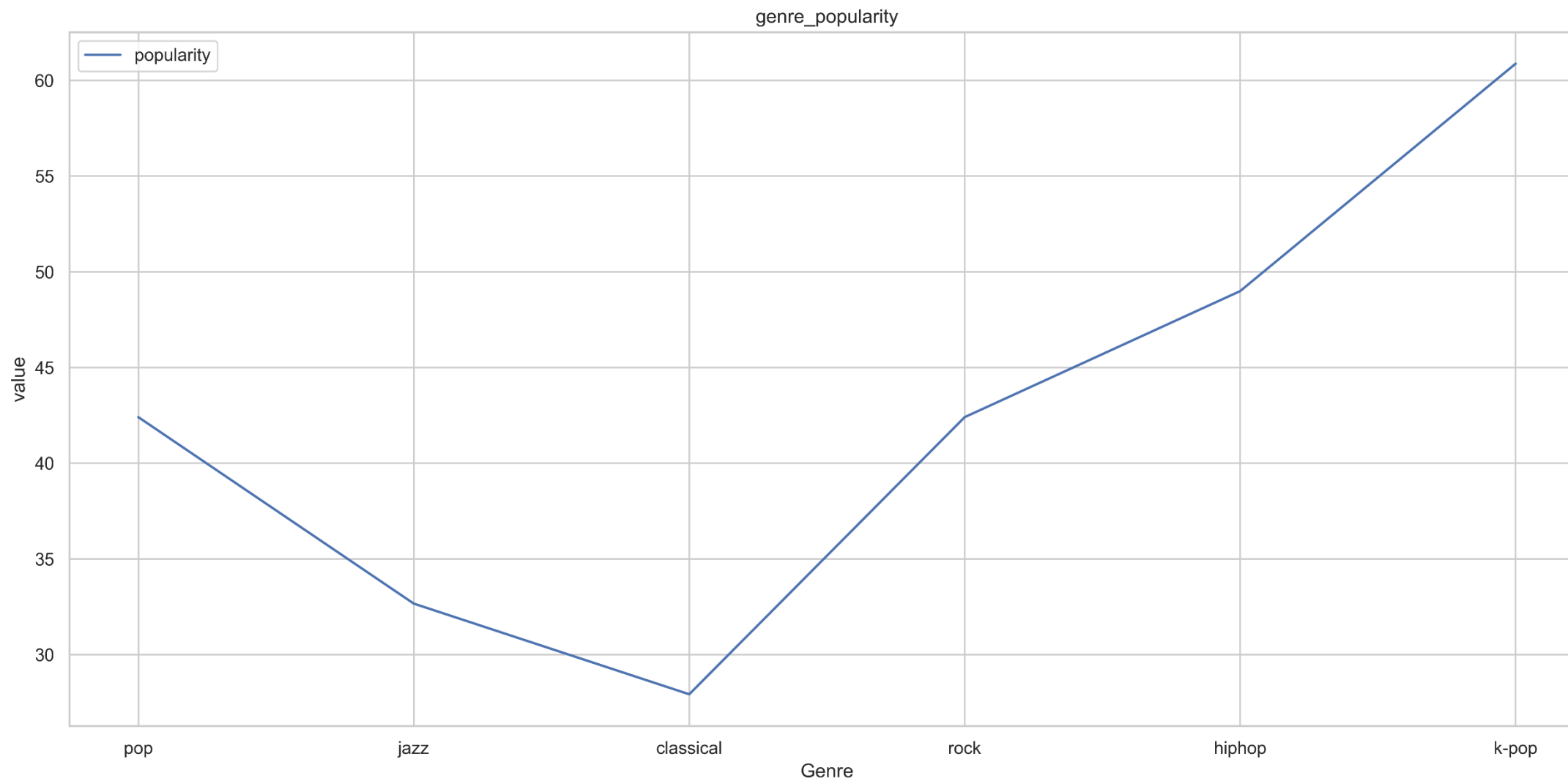
### 3. EDA

#### Genre Feature



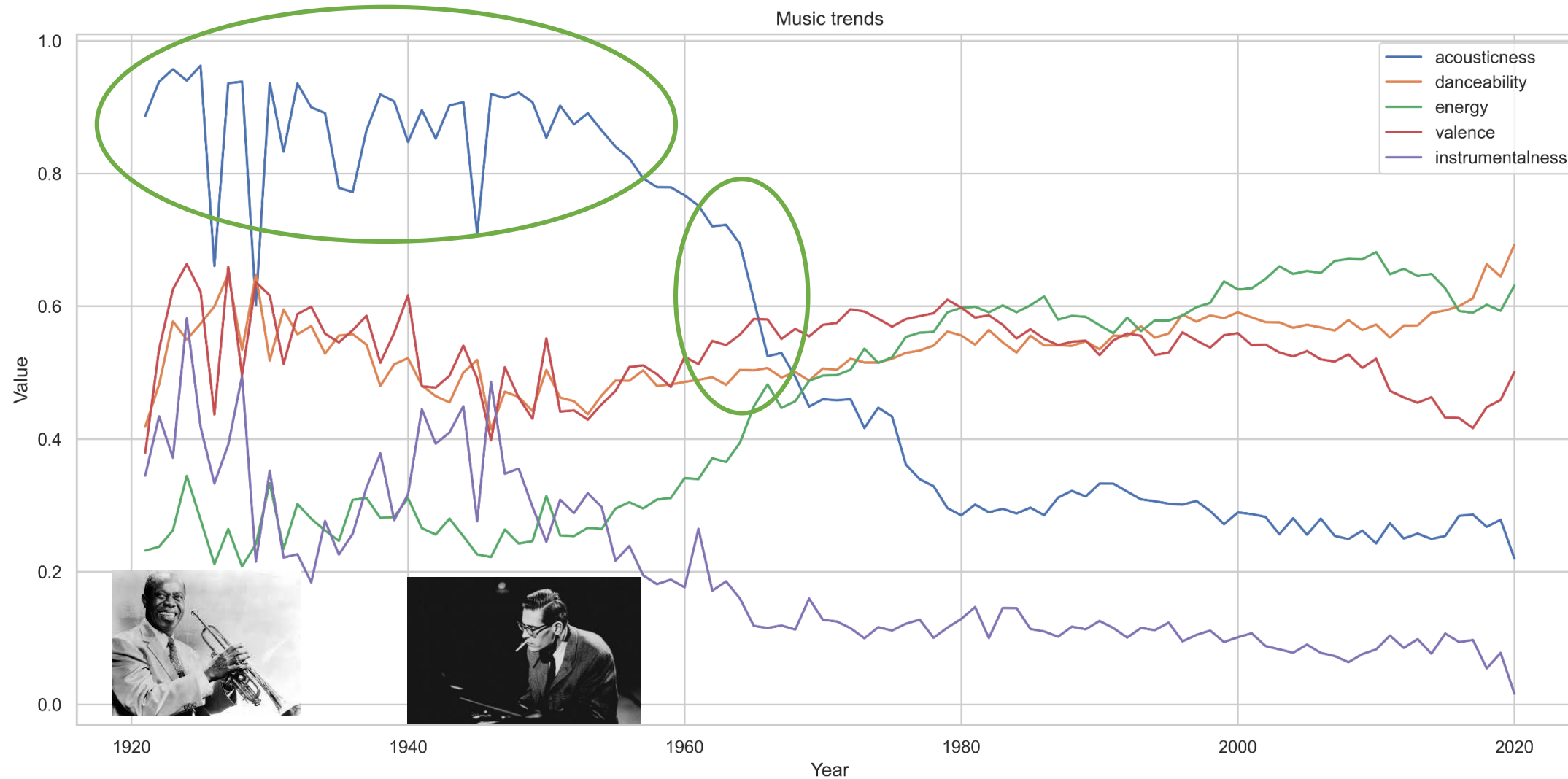
### 3. EDA

#### Genre Feature



### 3. EDA

#### Time Series





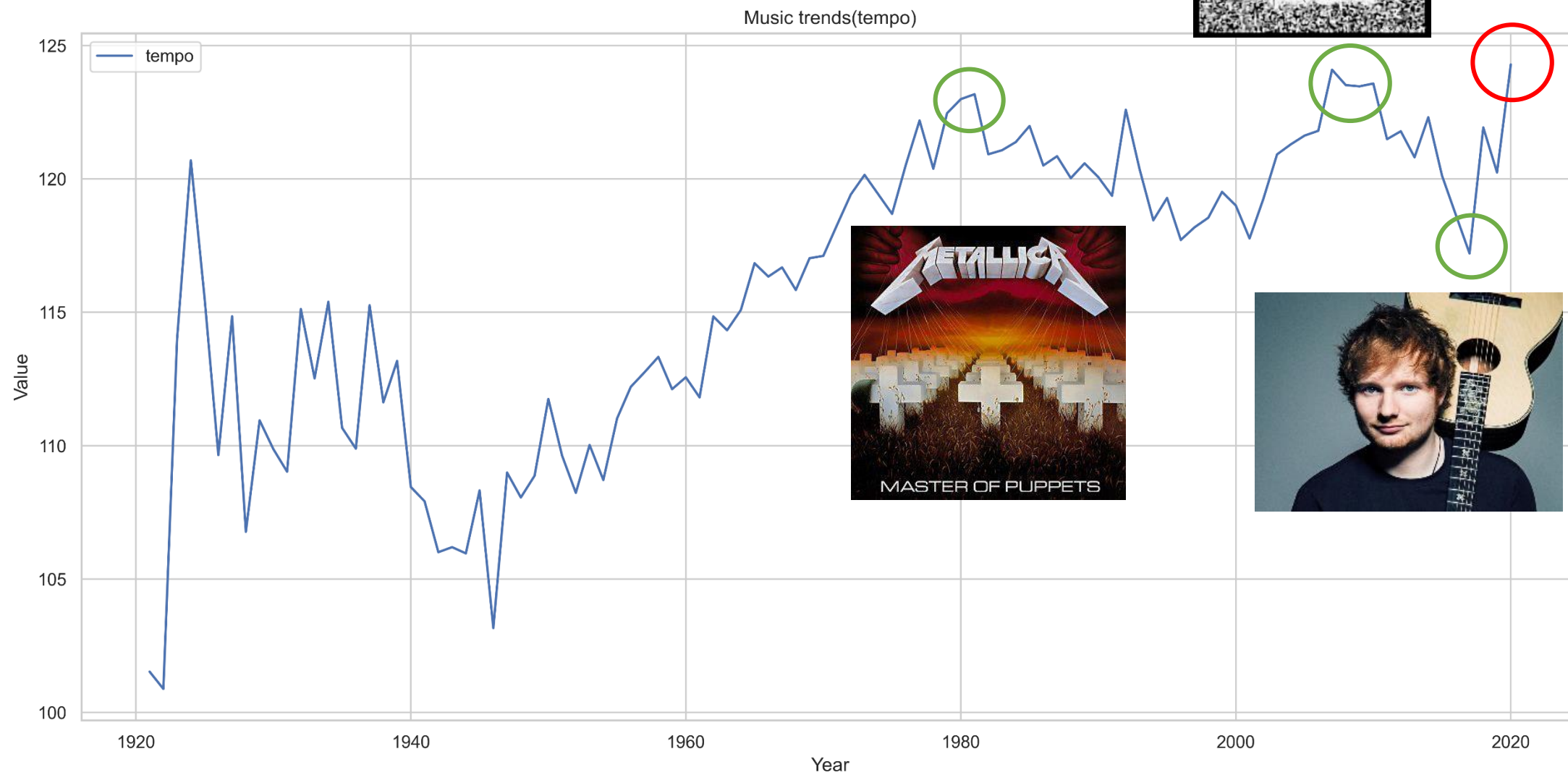
### 3. EDA

#### Time Series



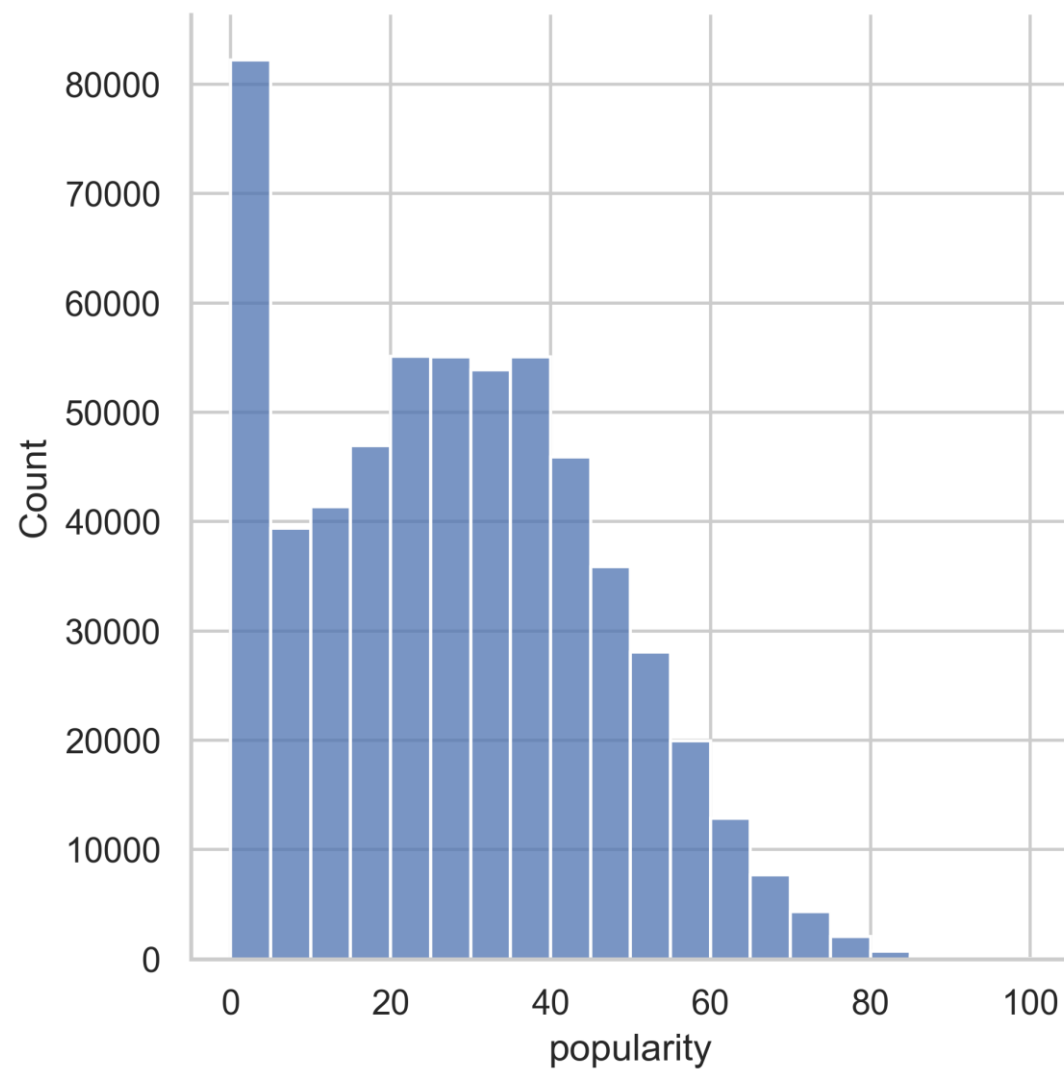
### 3. EDA

#### Time Series



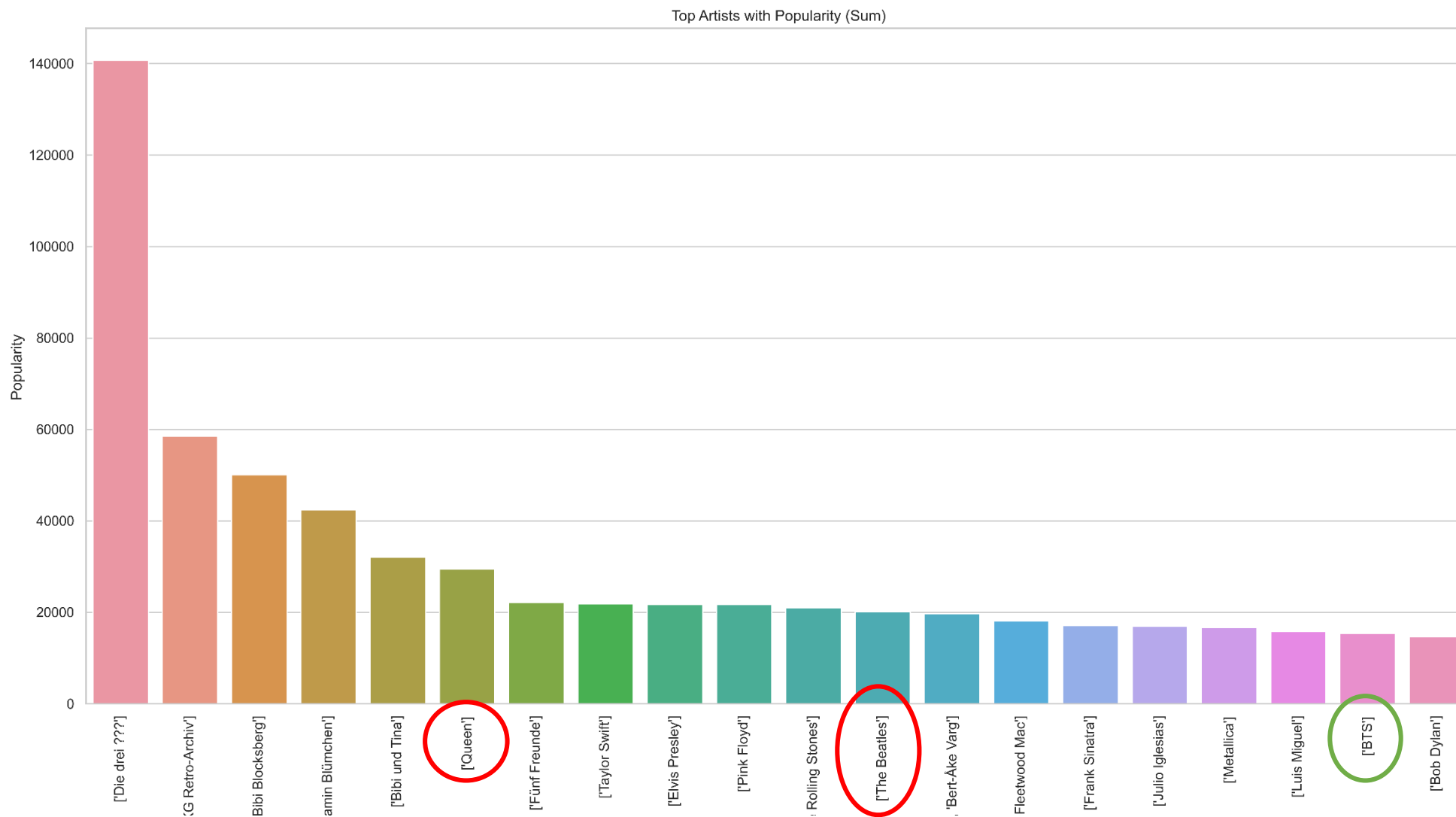
### 3. EDA

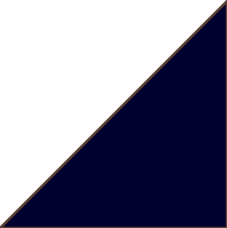
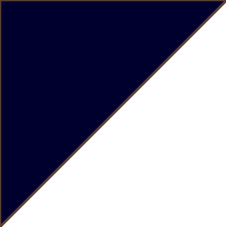
Artist



### 3. EDA

#### Artist





# 4. Outro

### 3. Outro

#### Conclusion

1. Spotify의 labeling이 실제로 유효하게 작동한다! (비법을 손에 넣었다!)
2. 최근 energy, loudness 한 음악이 대중적인 인지도를 얻는 추세다.
3. 반대로 acousticness, instrumental 한 음악은 대중적인 인지도를 얻지 못한다.
4. K-POP 위의 조건을 어느정도 충족하는 편이다. (세계 기준!)

**-> K-POP 트렌드를 많이 배워본다면 승산이 있지 않을까??**

### 3. Outro

---

#### Feedback

1. 하위 장르를 제대로 구분할 방법이 필요 (indie rock, indie pop, etc...)
2. 변인들이 만들어지는 원리 파악 (ex : danceability가 측정되는 구조)
3. Outlier, 주제와 거리가 먼 데이터 전처리 필요 (아동 애니메이션 주제가 등...)
4. mode, key를 활용한 음악적 접근의 EDA
5. \*사회 현상과 밀접하게 분석도 필요
6. \*10년 단위로 더 디테일하게 분석하는 것도 좋아보인다!

## 4. Outro

---

### Reference

- Dataset : <https://www.kaggle.com/yamaerenay/spotify-dataset-19212020-160k-tracks>
- Recommending music on Spotify with deep learning  
(<https://benanne.github.io/2014/08/05/spotify-cnns.html>)
- Spotify Web API (<https://developer.spotify.com/documentation/web-api/reference>)
- 스포티파이가 추천 맛집이 된 세 가지 이유 (<https://maily.so/musicdata/posts/171104>)





**Thank you**