Rock vs Hip-Hop

Gabriel Kremer, Sebastián Zapata Valencia, Alexander Echeverry

Problem Context

- Examine various data from different sources.
- Categorizing songs as either Rock or Hip-Hop without listening to them.
- Aims to:
  - Clean data
  - Make exploratory analysis
  - Increment dataset registers
  - Machine Learning model usage
  - Characteristics reduction

Data Overview

Attribute Description

1. ***Track Id:*** A unique identifier for every song.
2. ***Bit Rate:*** The bit rate of the audio file, which represents the amount of data processed per unit of time. Typically measured in bits per second (bps) and indicates the audio file's quality or compression level.
3. ***Comments:*** The number of comments or user-generated text responses associated with the song, often found on online music platforms or social media.
4. ***Composer:*** The name of the composer who created the music or wrote the song.
5. ***Date Created:*** The date when the song or audio file was originally created or uploaded.
6. ***Date Recorded:*** The date when the song was recorded, which may be different from the date it was created or uploaded.
7. ***Duration:*** The length of the song or audio file in terms of time, usually measured in seconds, minutes, or hours.
8. ***Favorites:*** The number of times users have marked the song as a favorite or liked it on a music platform.
9. ***Genre top:*** The primary or main genre classification of the song, indicating the style or category of music it belongs to.
10. ***Genres:*** A list of additional genres or subgenres that the song may be associated with, providing more detailed information about its musical style.

Attribute Description

11. ***Genres All:*** A comprehensive list of all genres and subgenres associated with the song, including both primary and secondary classifications.

12. ***Information:*** Additional information or metadata related to the song, which may include details about the artist, album, or other relevant information.

13. ***Interest:*** The level of interest or popularity of the song, often measured by metrics such as play count or user engagement.

14. ***Language Code:*** A code representing the language in which the song's lyrics or metadata are written, following language coding standards.

15. ***License:*** The type of license or legal terms associated with the song, indicating how it can be used, shared, or distributed.

16. ***Listens:*** The number of times the song has been listened to or streamed by users on a music platform.

17. ***Lyricist:*** The name of the lyricist or songwriter who wrote the lyrics for the song.

18. ***Number:*** A numerical identifier or track number within an album or playlist, used to order songs.

19. ***Publisher:*** The name of the publishing company or entity responsible for distributing or promoting the song.

20. ***Tags:*** Descriptive keywords or tags associated with the song, providing information about its content, mood, or themes.

Attribute Description

21. ***Title:*** The title or name of the song.
22. ***Acousticness:*** A measure of the acoustic characteristics of the song, indicating how much of the sound is generated by acoustic instruments (e.g., acoustic guitars, pianos) as opposed to electronic or synthesized sounds.
23. ***Danceability:*** A measure of the song's suitability for dancing, based on factors such as tempo, rhythm, and beat.
24. ***Energy:*** A measure of the song's energy level or intensity, often associated with its loudness and speed.
25. ***Instrumentalness:*** A measure of the song's instrumental nature, indicating the presence of vocals (or lack thereof) in the track.
26. ***Liveness:*** A measure of the song's perceived live performance quality, indicating the presence of audience sounds or live elements.
27. ***Speechiness:*** A measure of the song's speech-like elements, such as spoken words or vocal components that are not sung.
28. ***Tempo:*** The tempo of the song, representing its speed or beats per minute (BPM).
29. ***Valence:*** A measure of the song's mood or emotional positivity, with higher values indicating a more positive or joyful mood and lower values indicating a more negative or sad mood.

.csv file

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 17734 entries, 0 to 17733
Data columns (total 21 columns):
 #   Column         Non-Null Count  Dtype
---  ------         --------------  -----
 0   track_id       17734 non-null  int64
 1   bit_rate       17734 non-null  int64
 2   comments       17734 non-null  int64
 3   composer       166 non-null    object
 4   date_created   17734 non-null  object
 5   date_recorded  1898 non-null   object
 6   duration       17734 non-null  int64
 7   favorites      17734 non-null  int64
 8   genre_top      17734 non-null  object
 9   genres         17734 non-null  object
 10  genres_all     17734 non-null  object
 11  information    482 non-null    object
 12  interest       17734 non-null  int64
 13  language_code  4089 non-null   object
 14  license        17714 non-null  object
 15  listens        17734 non-null  int64
 16  lyricist       53 non-null     object
 17  number         17734 non-null  int64
 18  publisher      52 non-null     object
 19  tags           17734 non-null  object
 20  title          17734 non-null  object
dtypes: int64(8), object(13)
memory usage: 2.8+ MB
```

.json file

```
<class 'pandas.core.frame.DataFrame'>
Index: 13129 entries, 0 to 13128
Data columns (total 9 columns):
 #   Column            Non-Null Count  Dtype
---  ------            --------------  -----
 0   track_id          13129 non-null  int64
 1   acousticness      13129 non-null  float64
 2   danceability      13129 non-null  float64
 3   energy            13129 non-null  float64
 4   instrumentalness  13129 non-null  float64
 5   liveness          13129 non-null  float64
 6   speechiness       13129 non-null  float64
 7   tempo             13129 non-null  float64
 8   valence           13129 non-null  float64
dtypes: float64(8), int64(1)
memory usage: 1.0 MB
```

Merged Dataframe

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 4802 entries, 0 to 4801
Data columns (total 29 columns):
 #   Column            Non-Null Count  Dtype
---  ------            --------------  -----
 0   track_id          4802 non-null   int64
 1   bit_rate          4802 non-null   int64
 2   comments          4802 non-null   int64
 3   composer          106 non-null    object
 4   date_created      4802 non-null   object
 5   date_recorded     1234 non-null   object
 6   duration          4802 non-null   int64
 7   favorites         4802 non-null   int64
 8   genre_top         4802 non-null   object
 9   genres            4802 non-null   object
 10  genres_all        4802 non-null   object
 11  information       334 non-null    object
 12  interest          4802 non-null   int64
 13  language_code     2599 non-null   object
 14  license           4789 non-null   object
 15  listens           4802 non-null   int64
 16  lyricist          13 non-null     object
 17  number            4802 non-null   int64
 18  publisher         27 non-null     object
 19  tags              4802 non-null   object
 20  title             4802 non-null   object
 21  acousticness      4802 non-null   float64
 22  danceability      4802 non-null   float64
 23  energy            4802 non-null   float64
 24  instrumentalness  4802 non-null   float64
 25  liveness          4802 non-null   float64
 26  speechiness       4802 non-null   float64
 27  tempo             4802 non-null   float64
 28  valence           4802 non-null   float64
dtypes: float64(8), int64(8), object(13)
memory usage: 1.1+ MB
```

Attribute Removal

Nan Attributes

composer	4696
date_created	0
date_recorded	3568
duration	0
favorites	0
genre_top	0
genres	0
genres_all	0
information	4468
interest	0
language_code	2203
license	13
listens	0
lyricist	4789
number	0
publisher	4775

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 4802 entries, 0 to 4801
Data columns (total 23 columns):
 #   Column            Non-Null Count  Dtype
---  ------            --------------  -----
 0   track_id          4802 non-null   int64
 1   bit_rate          4802 non-null   int64
 2   comments          4802 non-null   int64
 3   date_created      4802 non-null   object
 4   duration          4802 non-null   int64
 5   favorites         4802 non-null   int64
 6   genre_top         4802 non-null   object
 7   genres            4802 non-null   object
 8   genres_all        4802 non-null   object
 9   interest          4802 non-null   int64
 10  license           4789 non-null   object
 11  listens           4802 non-null   int64
 12  number            4802 non-null   int64
 13  tags              4802 non-null   object
 14  title             4802 non-null   object
 15  acousticness      4802 non-null   float64
 16  danceability      4802 non-null   float64
 17  energy            4802 non-null   float64
 18  instrumentalness  4802 non-null   float64
 19  liveness          4802 non-null   float64
 20  speechiness       4802 non-null   float64
 21  tempo             4802 non-null   float64
 22  valence           4802 non-null   float64
dtypes: float64(8), int64(8), object(7)
memory usage: 863.0+ KB
```

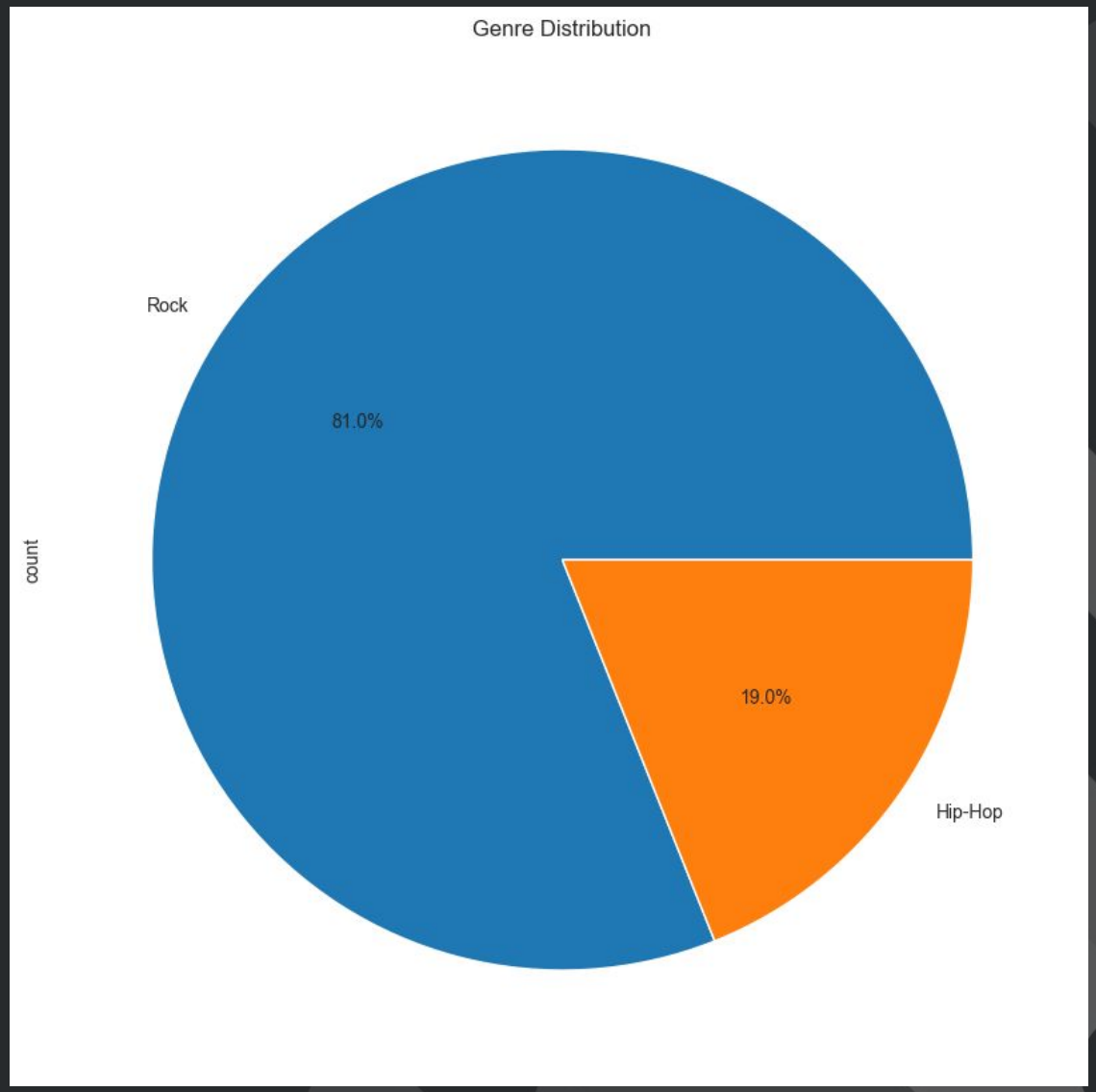
Irrelevant Attributes

- Track id?
- Bit rate
- Comments
- Date created
- Favorites
- Genres
- Genres All
- License
- Number

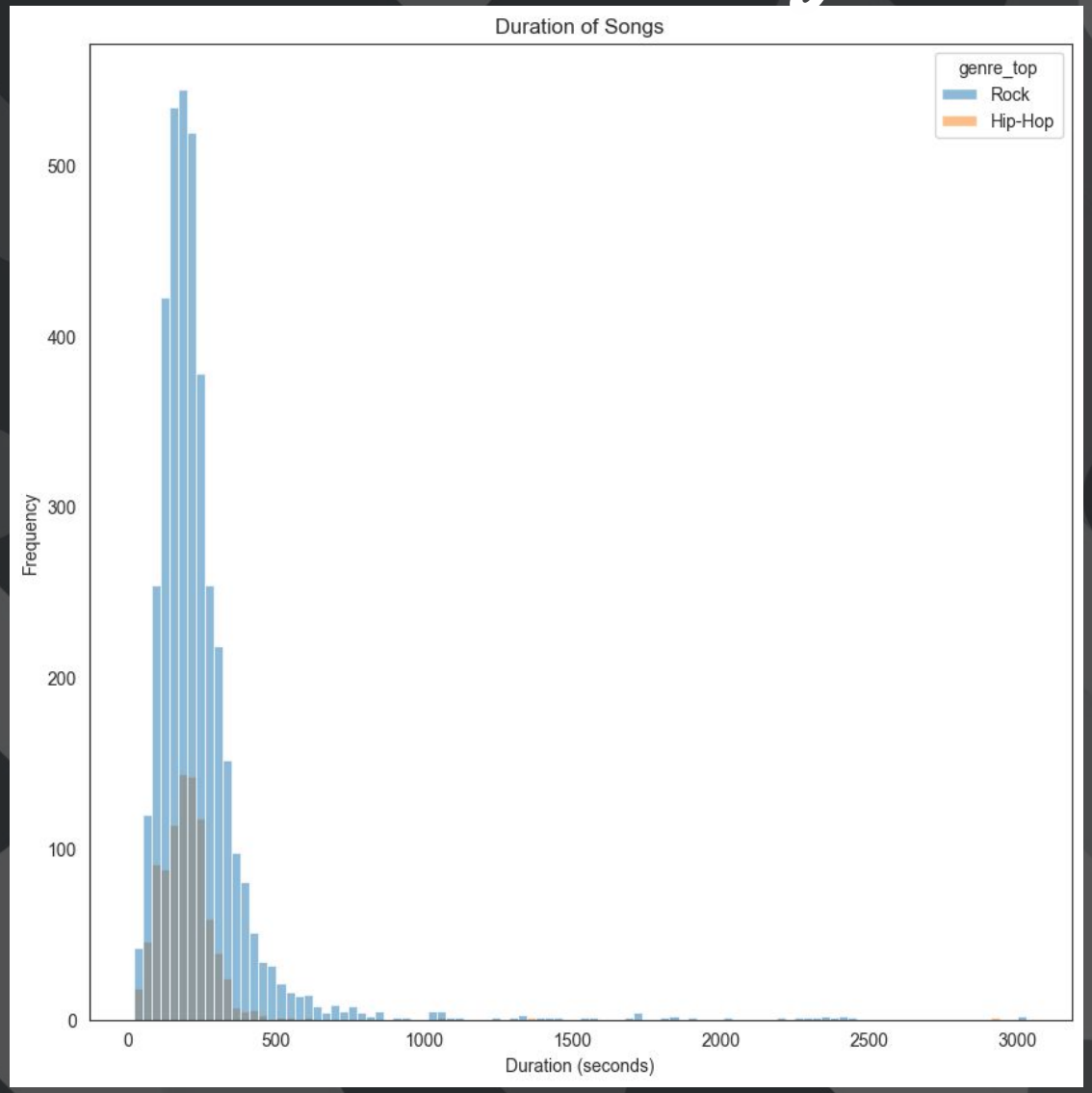
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 4802 entries, 0 to 4801
Data columns (total 15 columns):
 #   Column            Non-Null Count  Dtype
---  ------            --------------  -----
 0   track_id          4802 non-null   int64
 1   duration          4802 non-null   int64
 2   genre_top         4802 non-null   object
 3   interest          4802 non-null   int64
 4   listens           4802 non-null   int64
 5   tags              4802 non-null   object
 6   title             4802 non-null   object
 7   acousticness      4802 non-null   float64
 8   danceability      4802 non-null   float64
 9   energy            4802 non-null   float64
 10  instrumentalness  4802 non-null   float64
 11  liveness          4802 non-null   float64
 12  speechiness       4802 non-null   float64
 13  tempo             4802 non-null   float64
 14  valence           4802 non-null   float64
dtypes: float64(8), int64(4), object(3)
memory usage: 562.9+ KB
```

Univariate Analysis

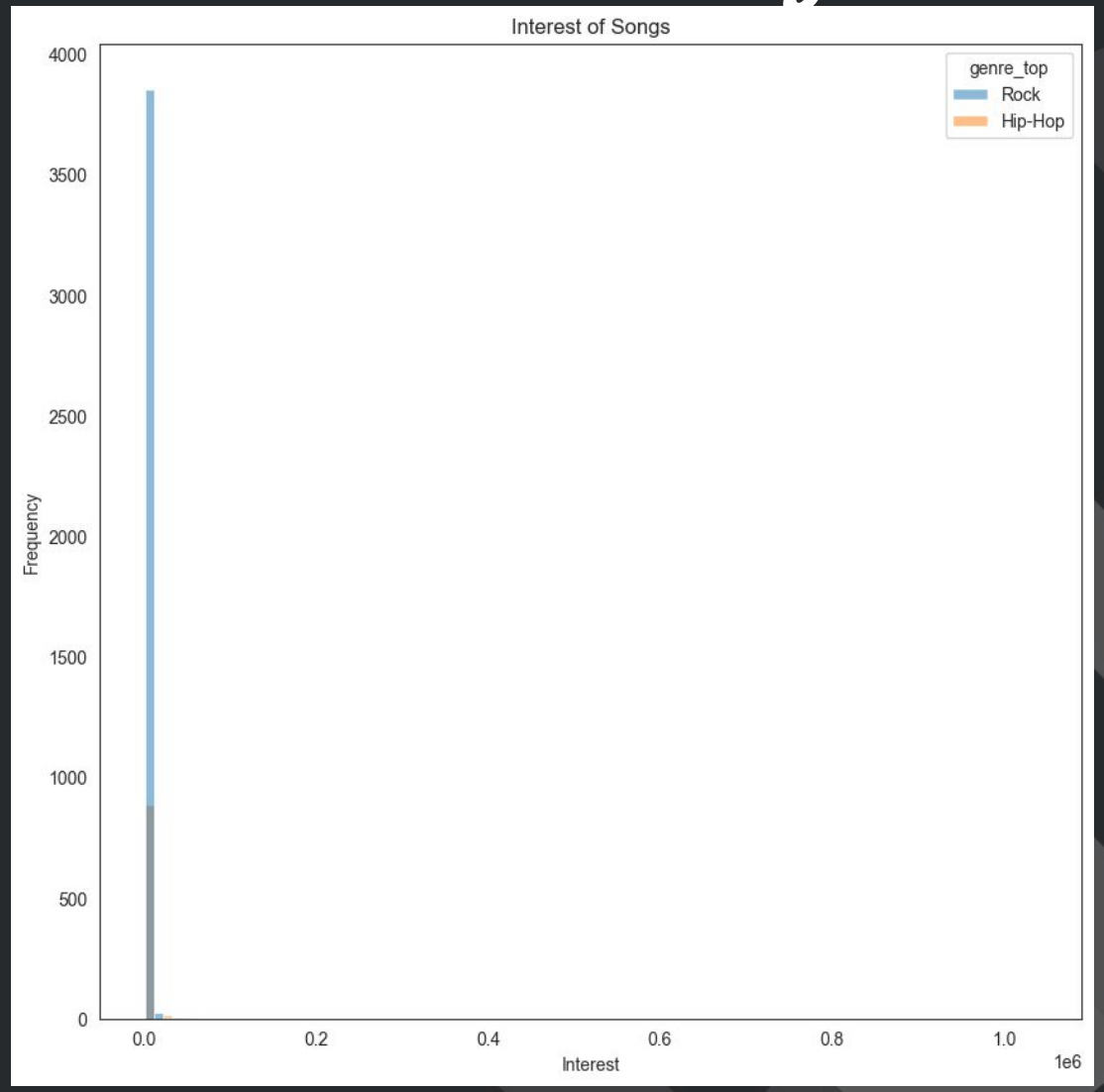
Genre Distribution



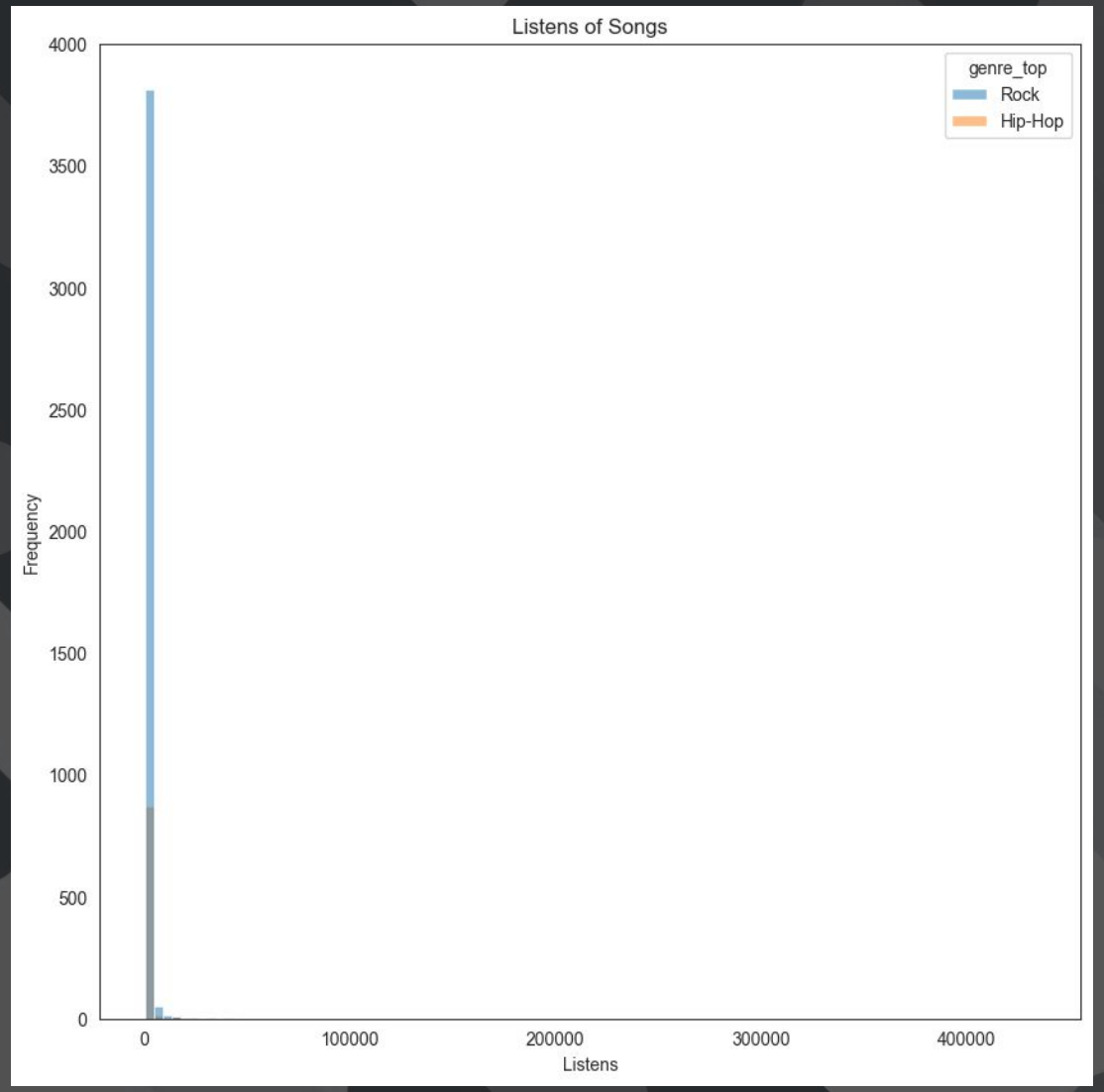
Duration Histogram



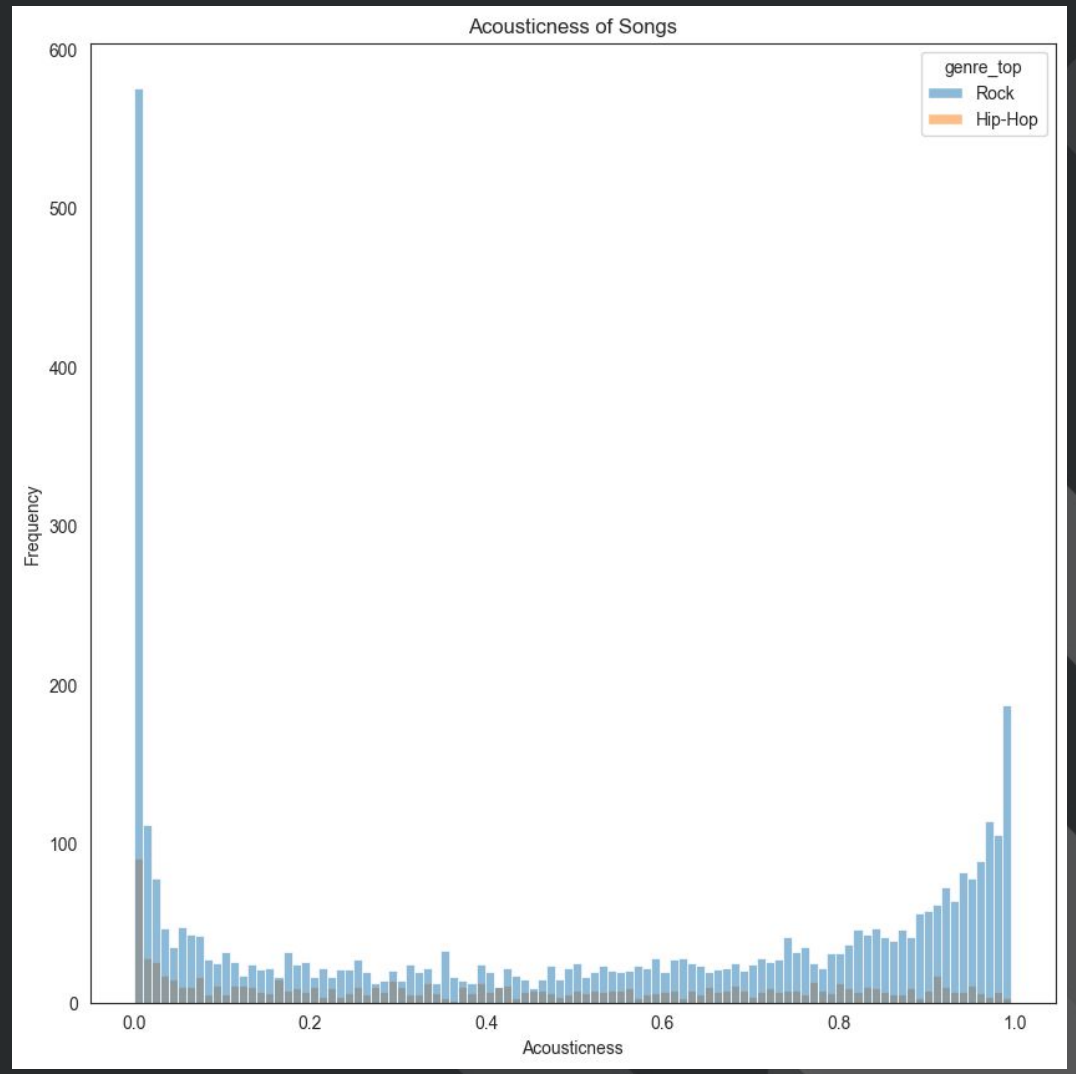
Interest Histogram



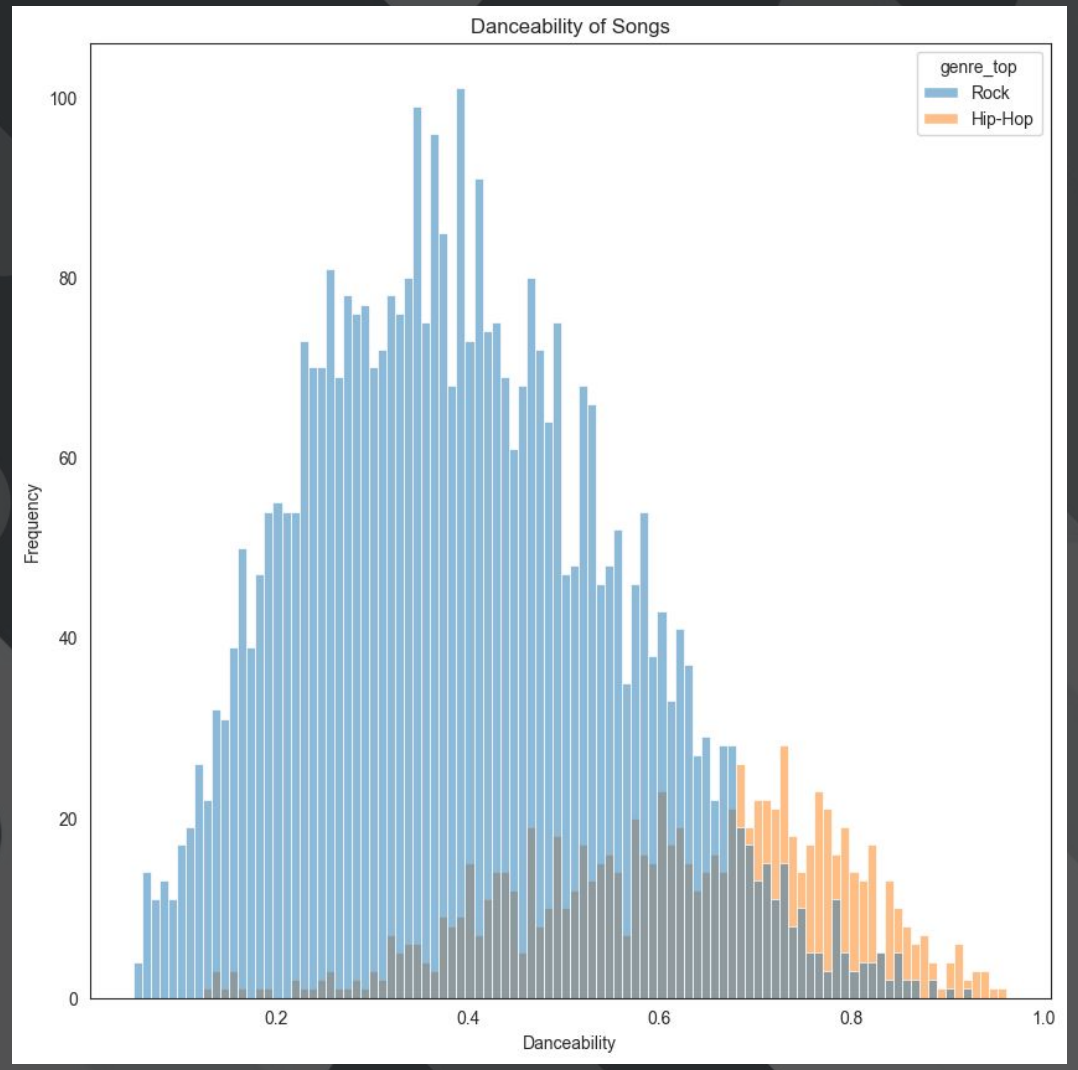
Listens Histogram



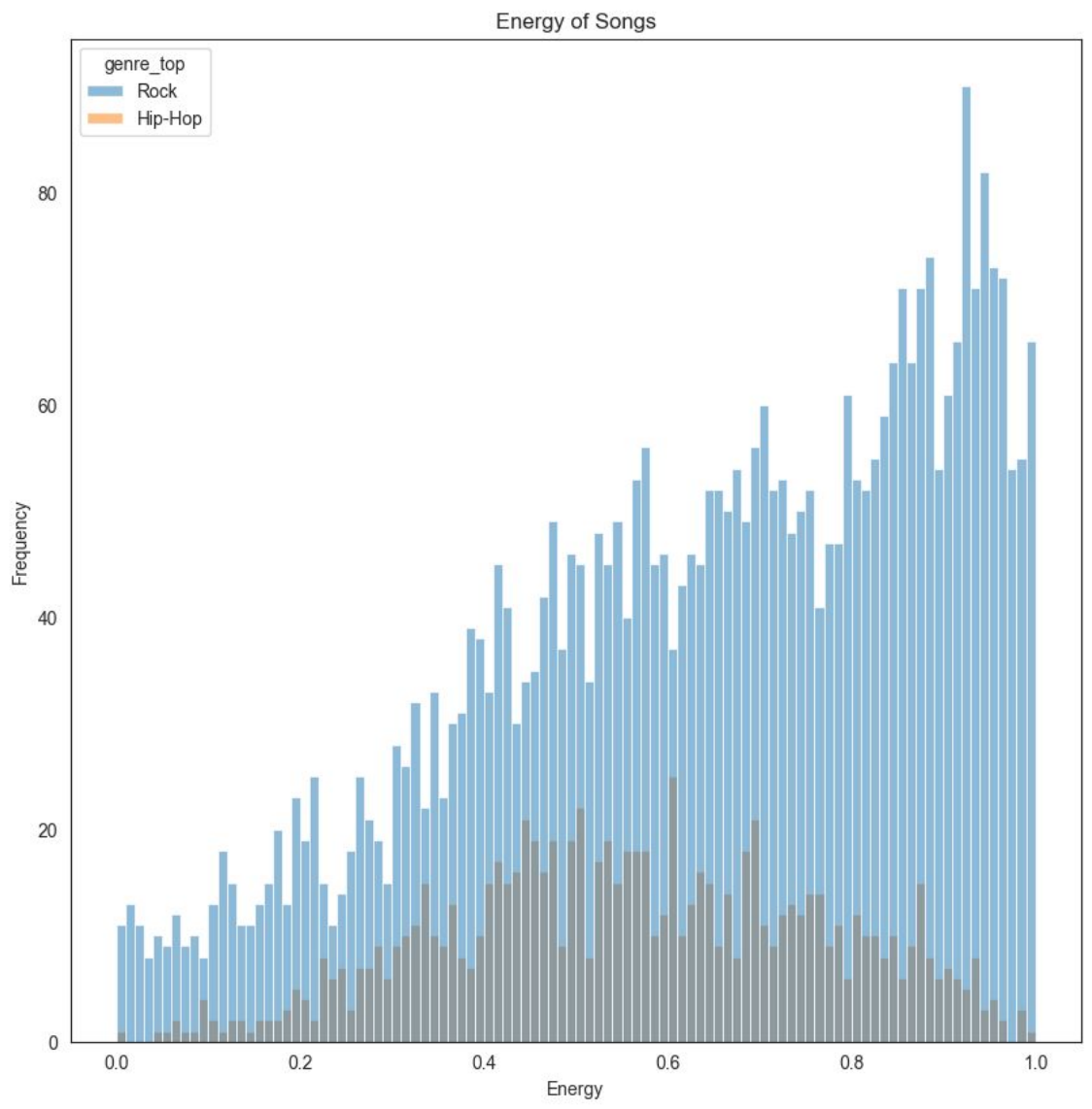
Acousticness Histogram



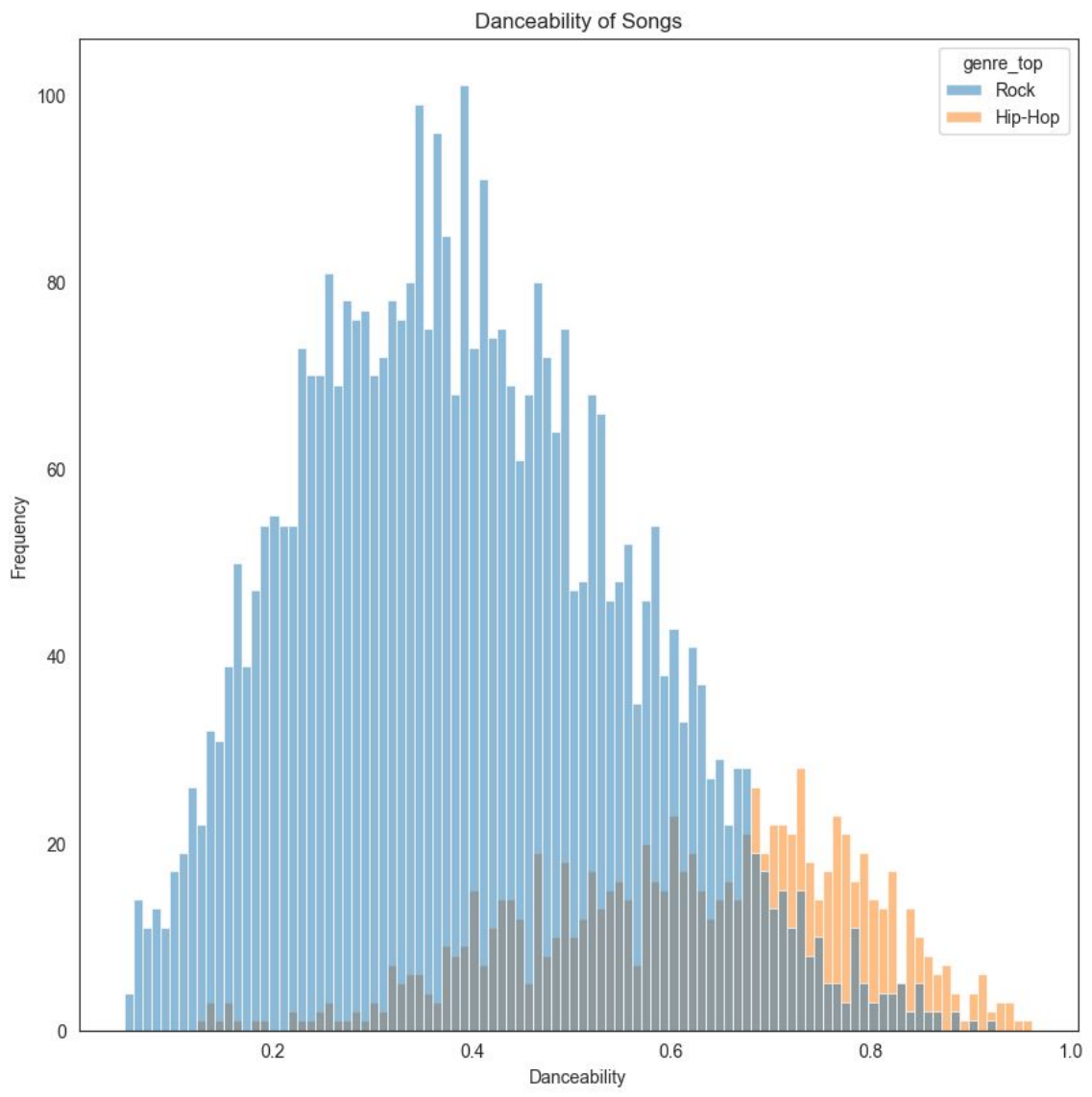
Danceability Histogram



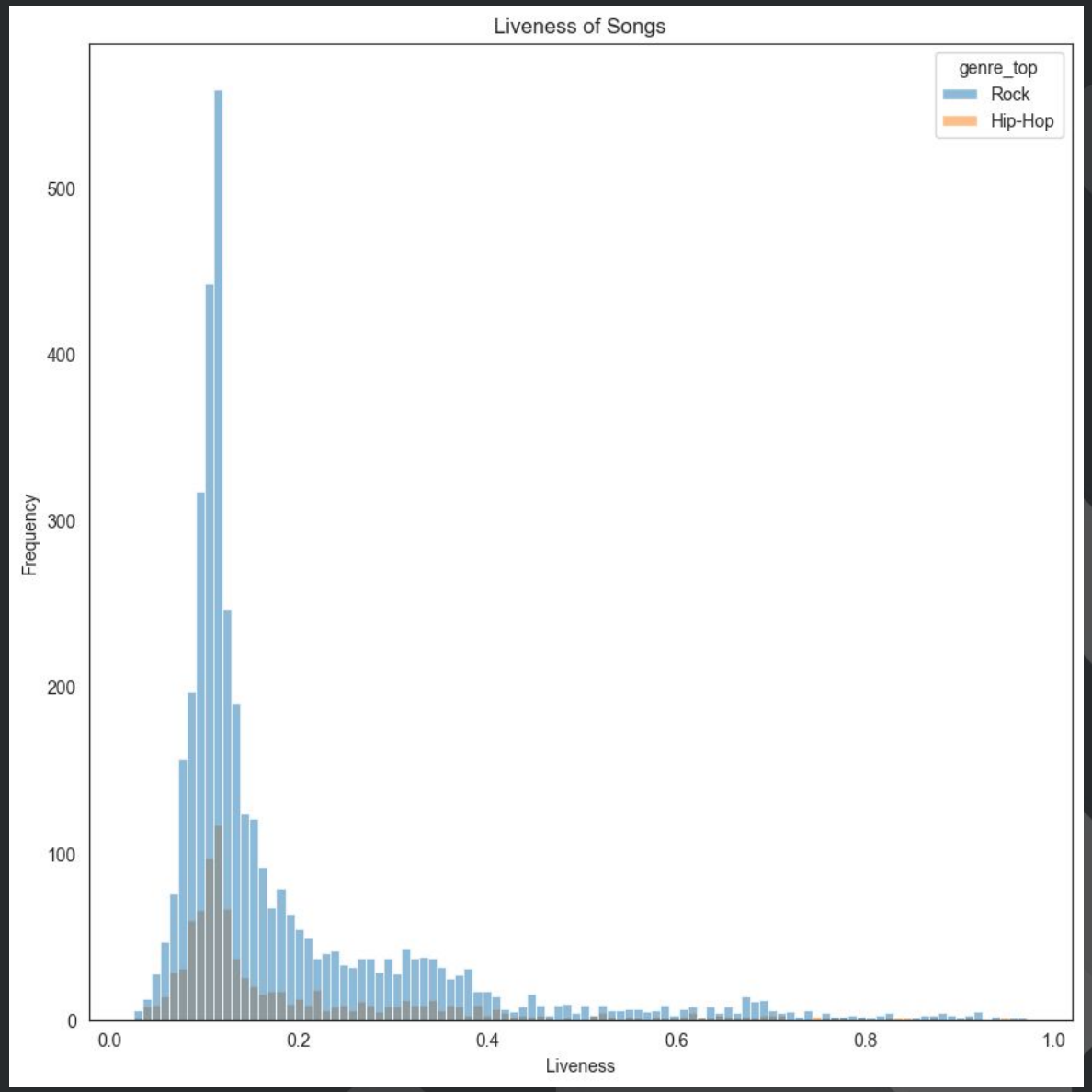
Energy Histogram



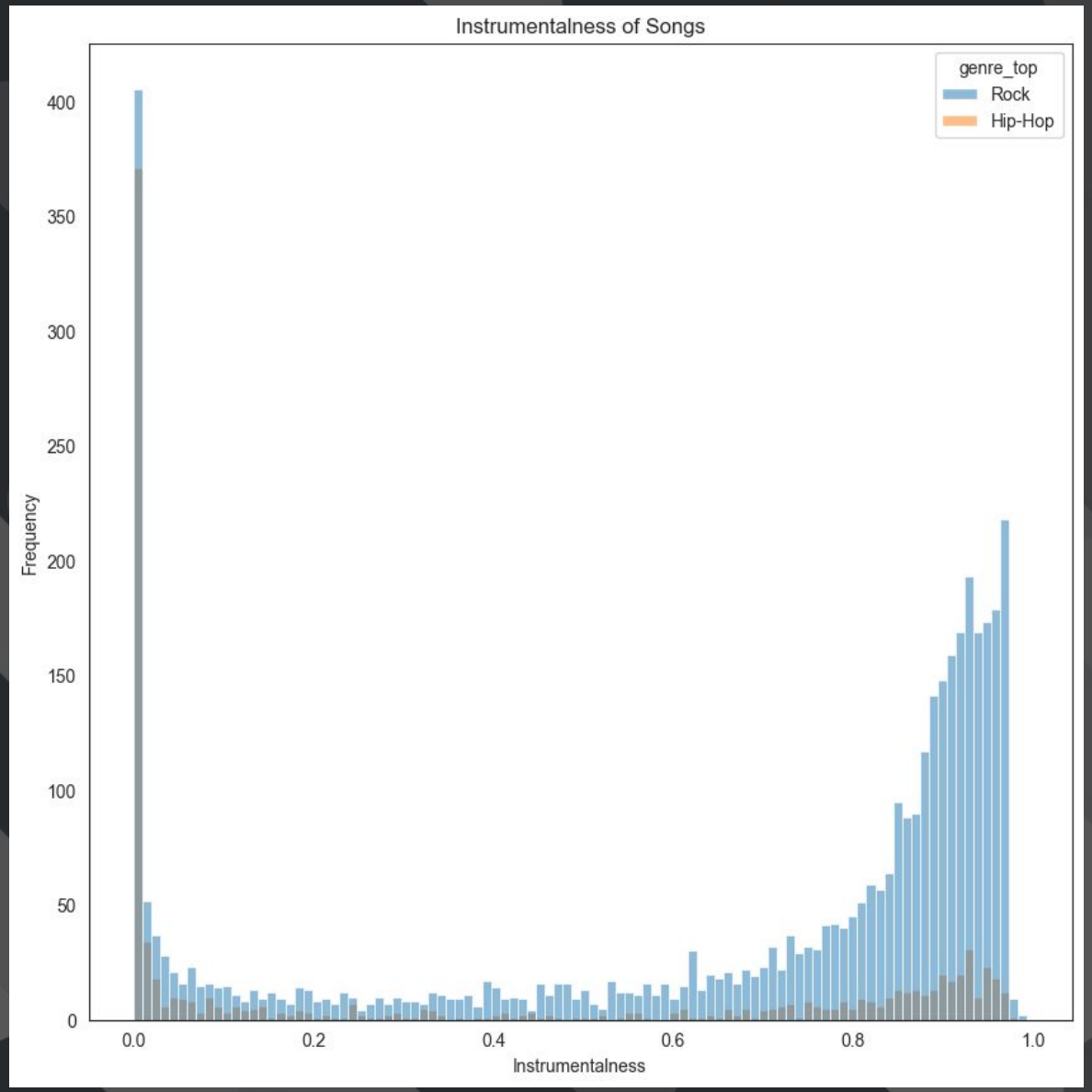
Danceability Histogram



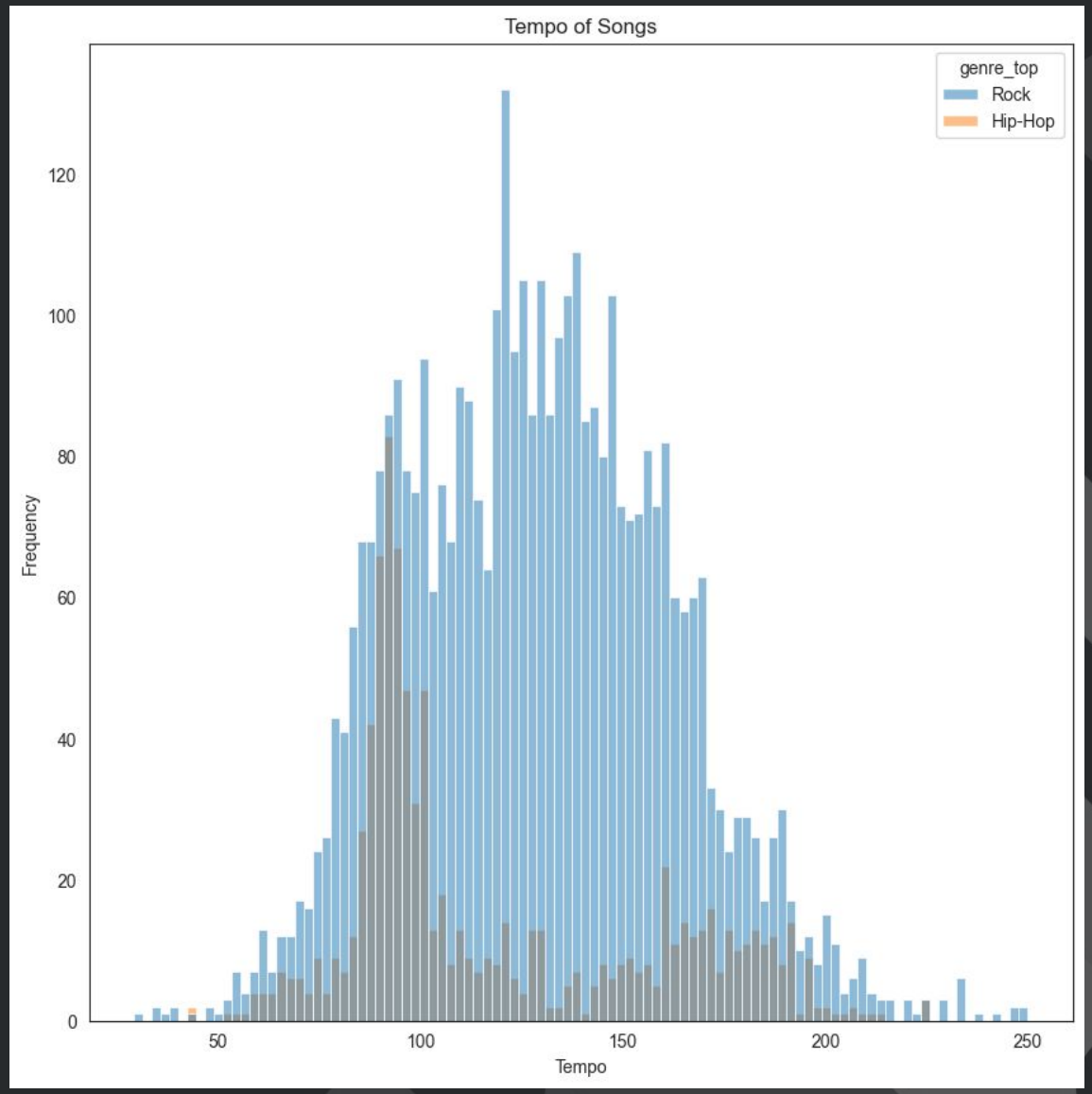
Liveness Histogram



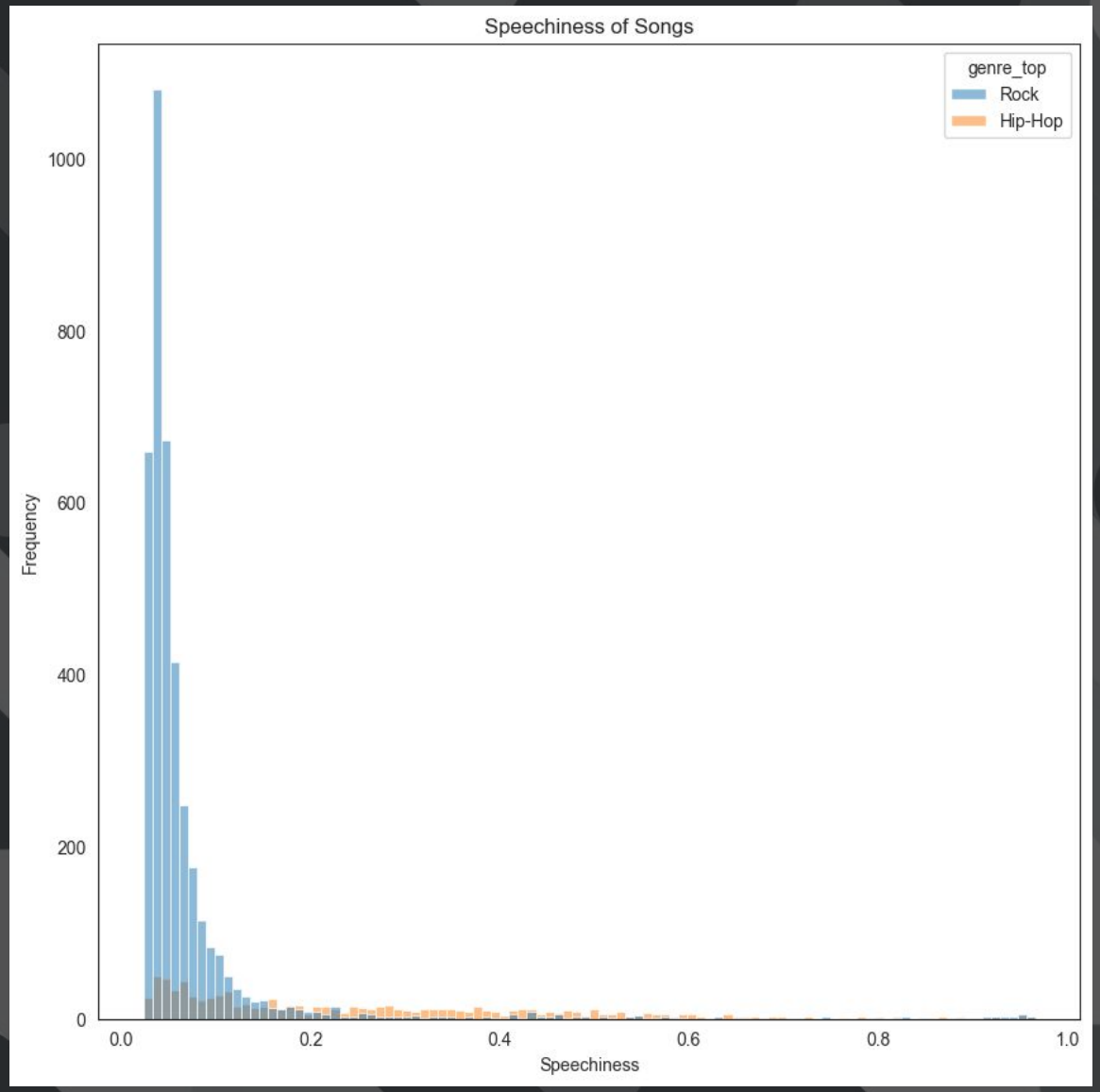
Instrumentalness Histogram



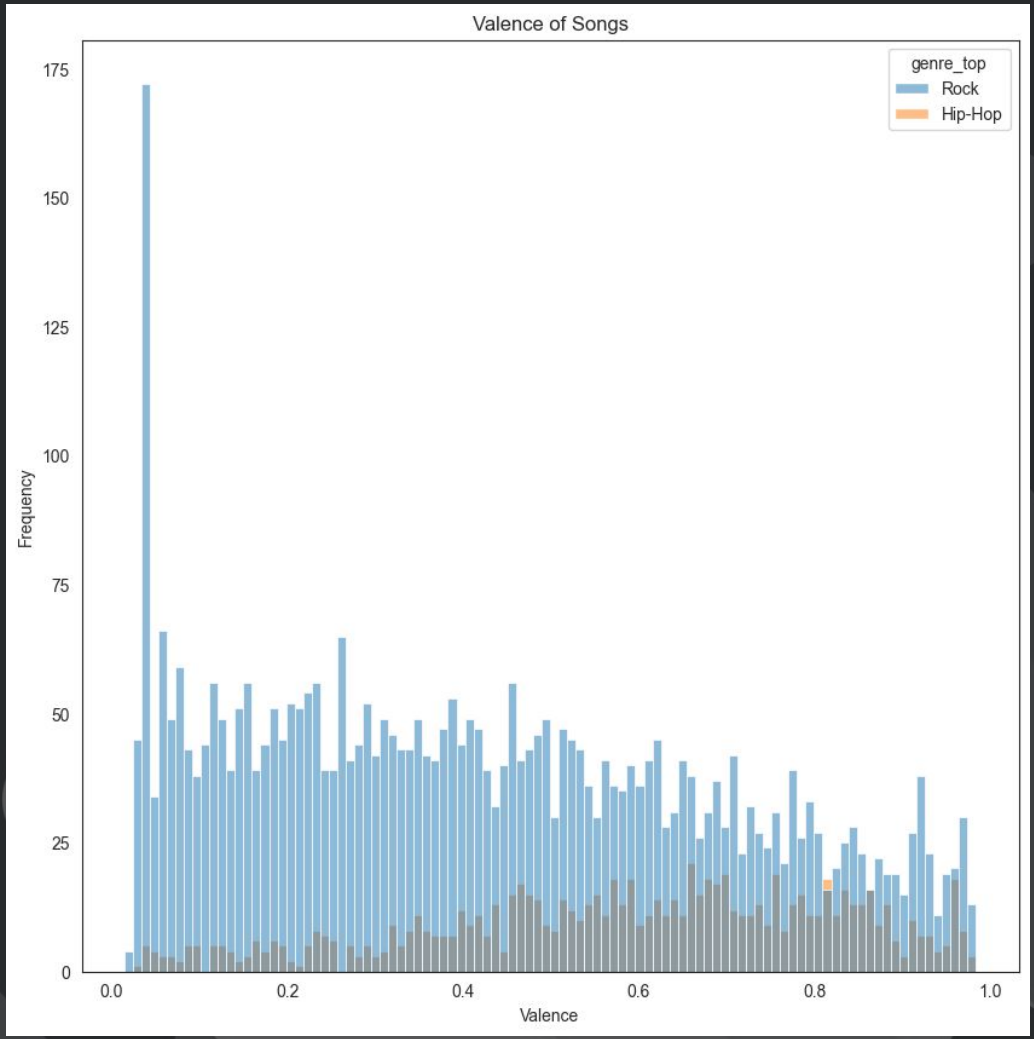
Tempo Histogram



Speechiness Histogram

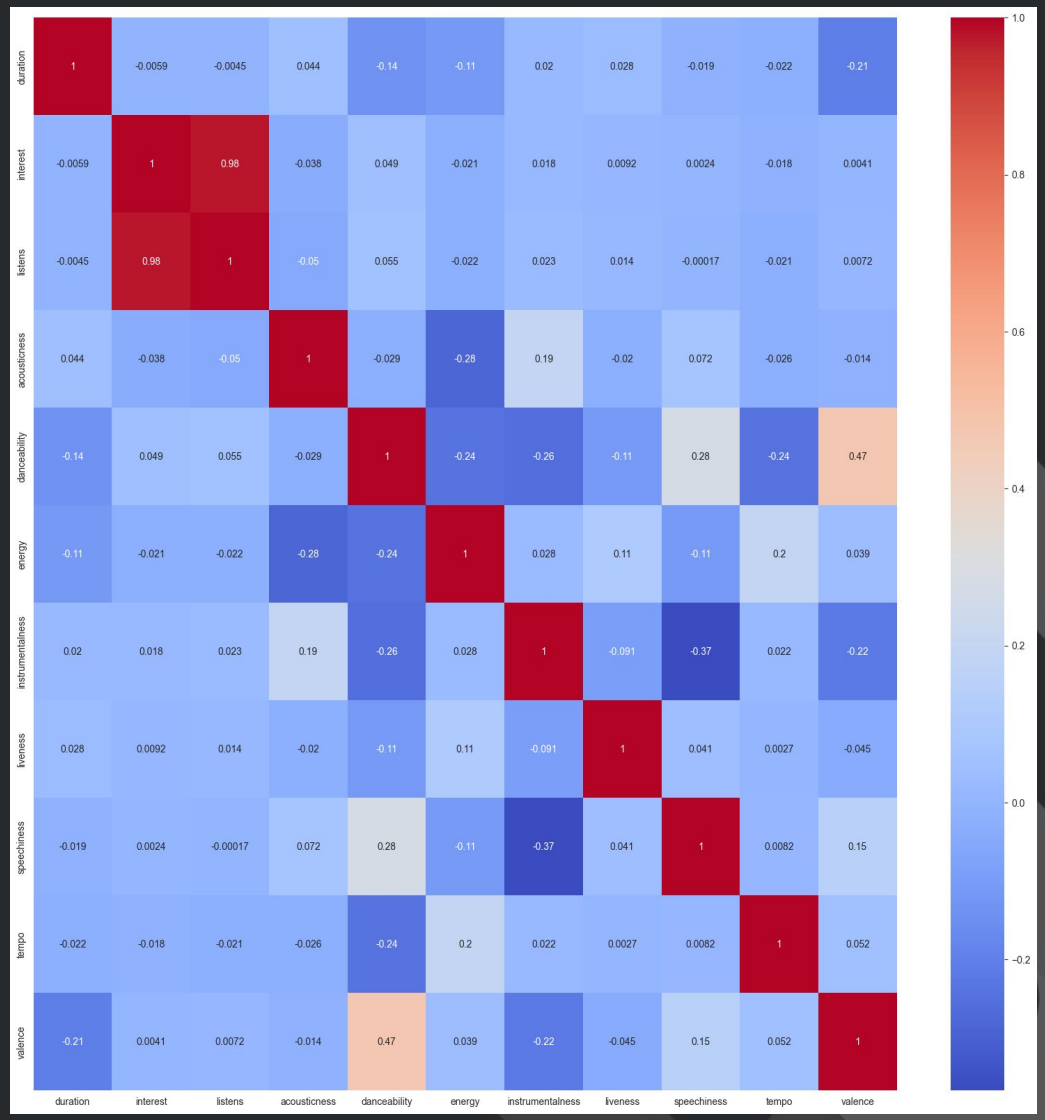


Valence Histogram

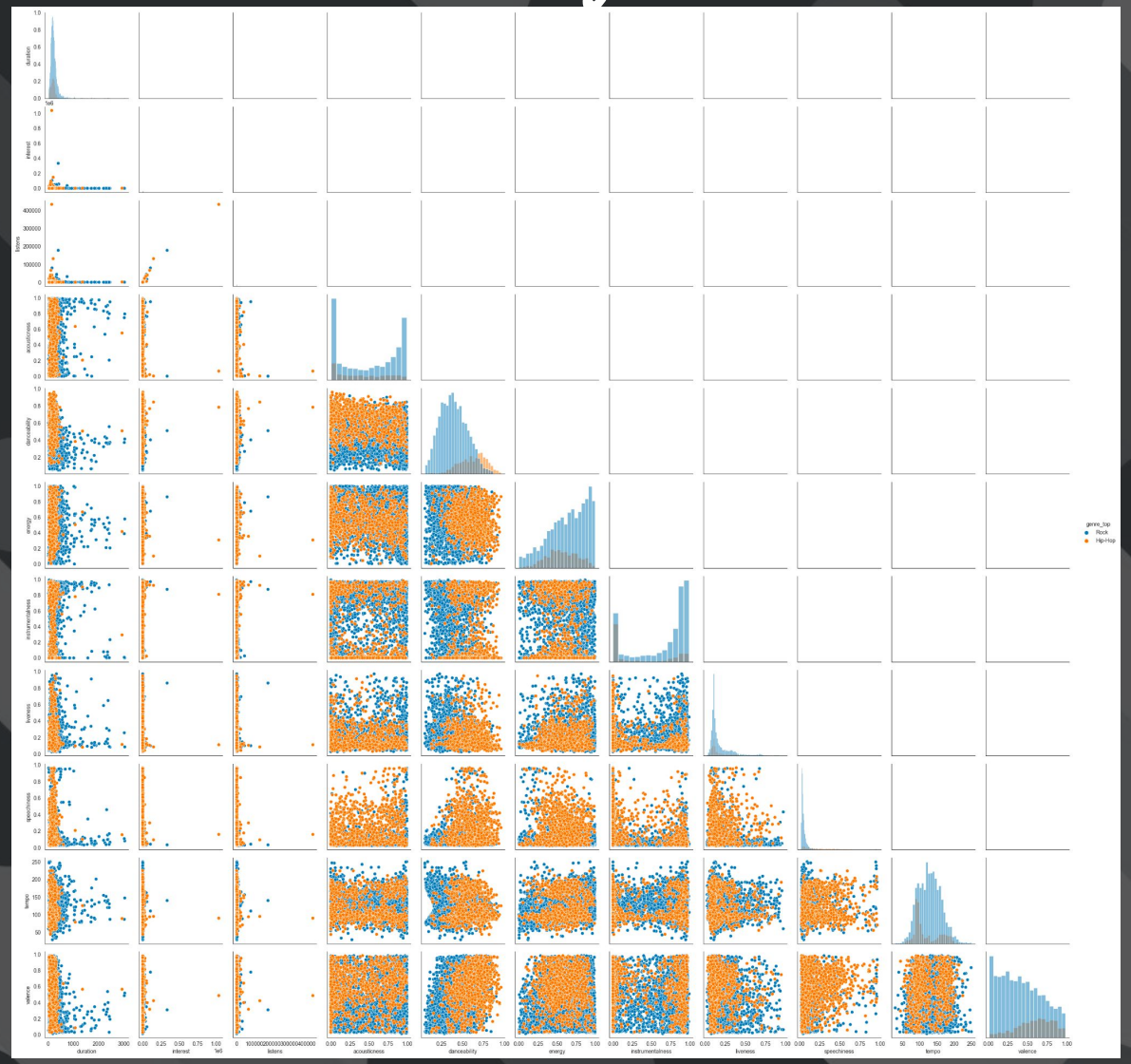


Bivariate Analysis

Correlation matrix



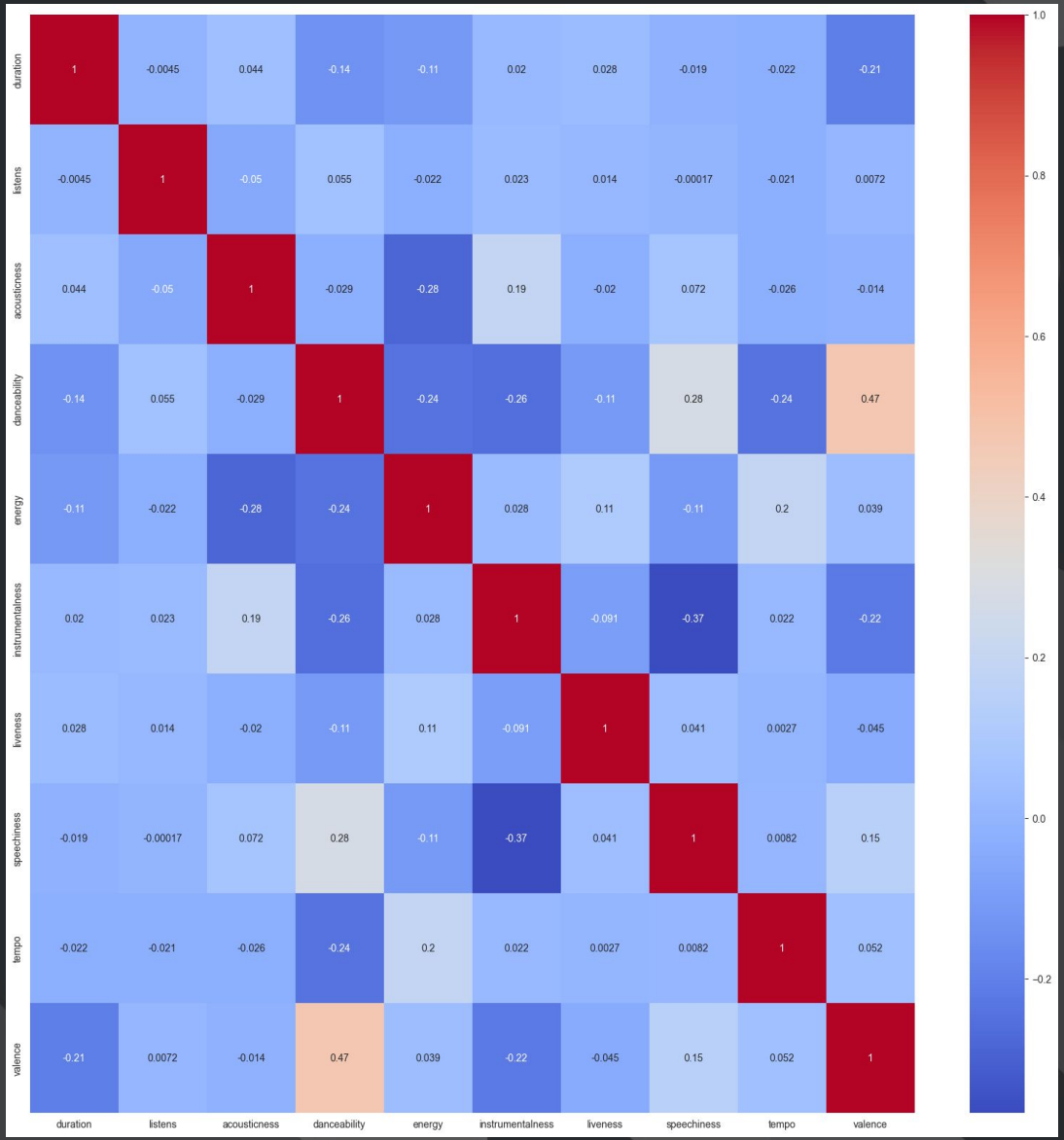
Scatterplots



Correlation between "Interests" and "Listens"

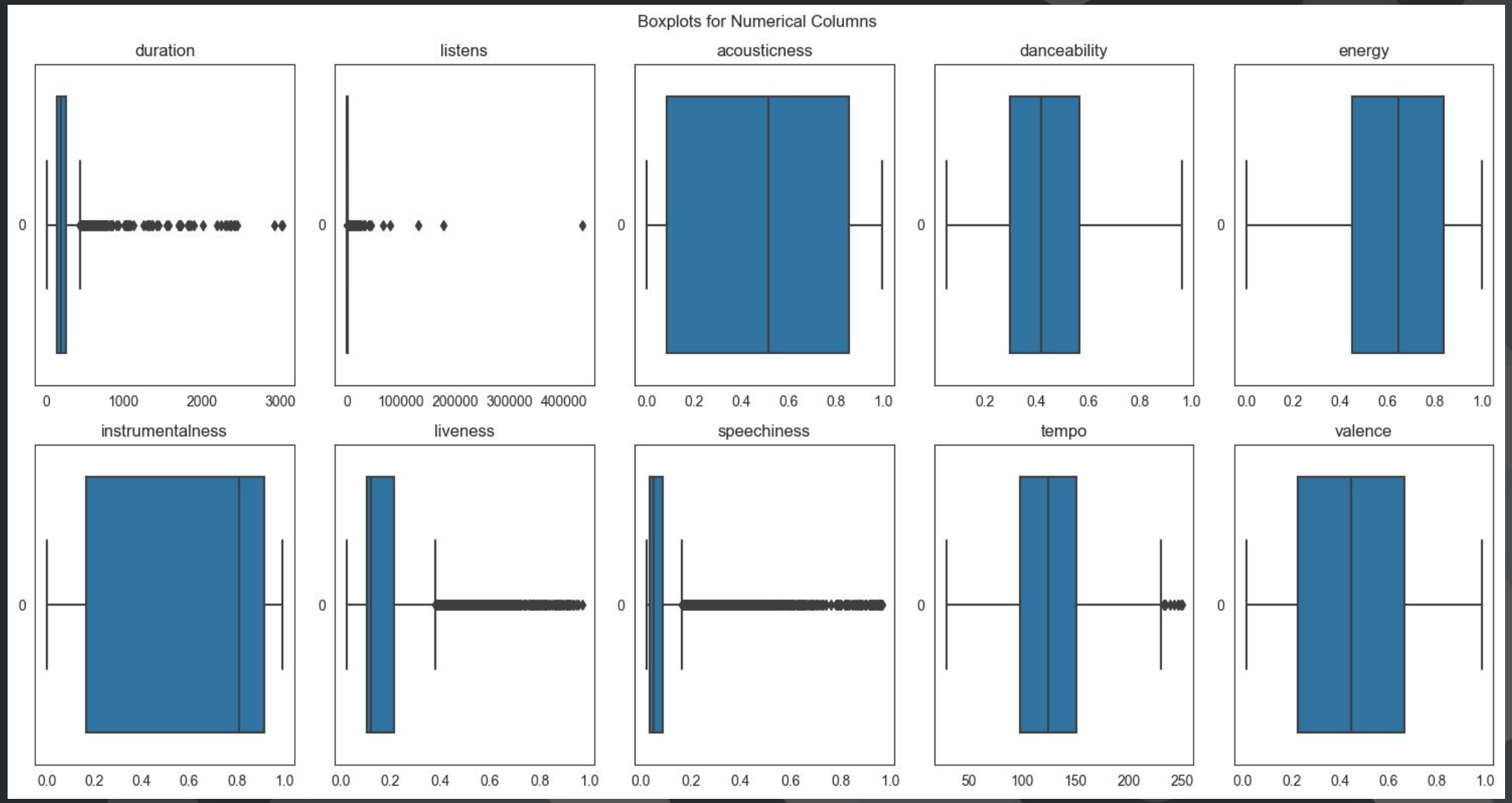
1	-0.0059	-0.0045	0.044
-0.0059	1	0.98	-0.038
-0.0045	0.98	1	-0.05

Corrected correlation matrix



Data Cleaning

Boxplots


Boxplots for Numerical Columns
duration
listens
acousticness
danceability
energy
instrumentalness
liveness
speechiness
tempo
valence
0
0 1000 2000 3000
0 100000 200000 300000 400000
0.0 0.2 0.4 0.6 0.8 1.0
0.2 0.4 0.6 0.8 1.0
50 100 150 200 250

Data Collection Strategies

Echo Nest API

Spotify API

Apple Music API

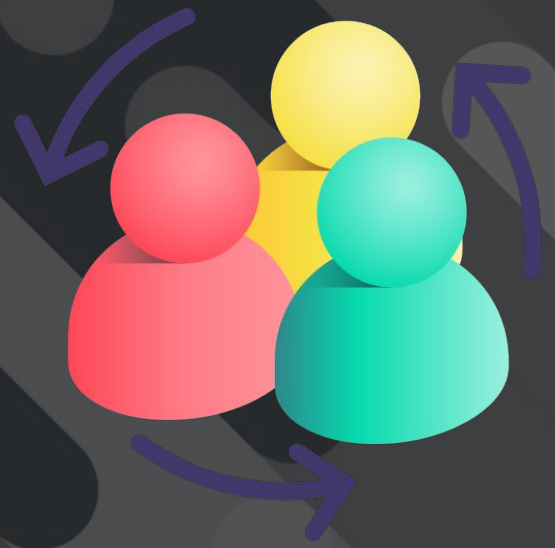
Artificial Intelligence Impact

Global Impact

- Technological Competitiveness
- National Security

Social Impact

- Labor Displacement
- Equity and Bias



Environmental Impact

- Energy Efficiency
- Environmental Monitoring
- E-waste products

Economical Impact

- Economic Growth
- Investment and Regulation
- New Business Models

END