DRAFT - CSC 454 Final Project

Dr. Christer Karlsson

CSC 454 - Data Mining Theory

February 18, 2022

Outlier Detection in Detecting Misclassified Music

There seems to be some logical relation between genre and the physical properties of music like tempo, danceability, tone, etc. Apart from that, there's instances where a certain piece of music has been classified as belonging to one genre but feel like they are outliers in that genre.

This document provides a rough draft of how Outlier Detection can be used to identify music that stands out of its genre - music that blends the lines between genres - and attempt to reclassify the music and provide a logical association between genre and other classifiers like danceability, tempo and so on.

Summary:

- Documentation of Dataset
- Hypothesis
- How the Study will be conducted

Documentation of Dataset

My initial choice for the dataset was my own personal music listening history on the music streaming app Spotify. This data contained the following attributes:

- trackName = The name of the track streamed
- artistName = The name of the artist that produced the track
- endTime = The time when I ended the track
- msPlayed = The total time I played the track for

This however was not enough and more attributes like tempo, major, danceability, etc were needed to. Since my music history mostly contained music that was produced more recently, valid data corresponding to the above mentioned attributes was not available.

Therefore the main dataset was changed to something similar, obtained from kaggle.com. This new chosen data was obtained from this open source project - Music Genre Classification - and contains the following attributes

1) **Class -** The genre
2) **Artist Name -** String attribute containing the name of the artist
3) **Track Name -** Name of the track
4) **Popularity -** How popular the song is (metric between 0 and 100)
5) **Danceability -** How easy it is to dance to the trace (metric between 0 and 1)

6) **Energy -** How lively a song is (metric between 0 and 1)

7) **Key -** The key in which the song is (metric between 0 to 11 corresponding to C - B)

8) **Loudness -** How loud the music is (metric in dB)

9) **Mode -** Corresponding to the musical definition of mode (binary)

10) **Speechiness -** Measure of the the presence of spoken words in a track (0 to 1)

11) **Acousticness -** Measure of how acoustic a track is (metric between 0 and 1)

12) **Instrumentalness -** Measure of how much of a track is just instrumentals (0 to 1)

13) **Liveness -** Measure of reverberation time (between 0 and 1)

14) **Valence -** Measure from 0 - 1 describing the musical positiveness of a track.

15) **Tempo -** Beats per minute of a track

16) **Duration -** Time duration of the track in ms

17) **Time Signature -** The musical time signature (3/4 or 4/4)

The **training dataset** contains **17,996 entries** with **17 attributes**, while the testing dataset contains **7,713 entries with 17 attributes** (show above). The datasets can be found at this link - datasets.

Ethical and Privacy Concerns :- The dataset was obtained from an open source data analysis project from **kaggle.com**

Figures and Visualization of Data

Figures 1 - 4 show Box Plots of the data corresponding to attributes that are easily perceived aurally or physically.

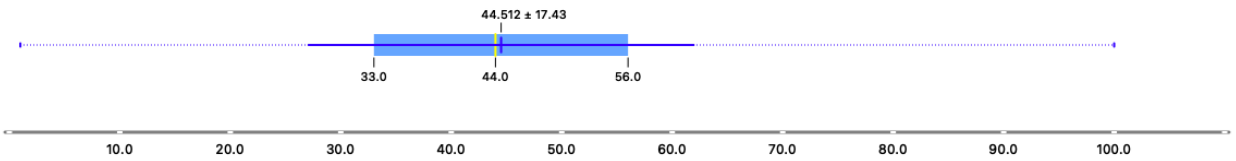

Figure 1: Popularity

From Figure 1 and Figure 2, we can infer that most of the data chosen is not super popular and includes a wide selection with high outliers.

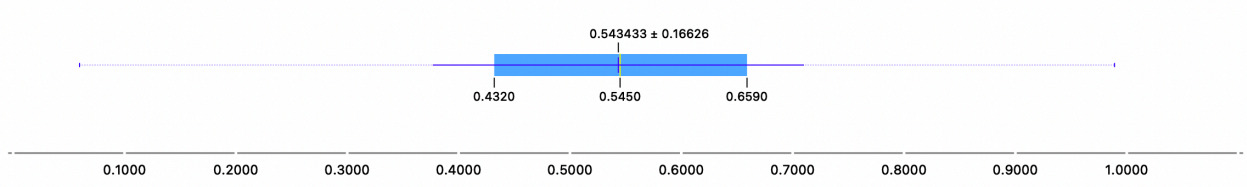

Figure 2: Danceability

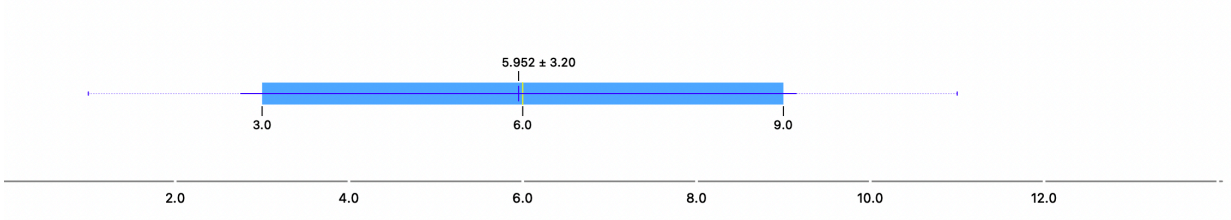

Figure 3: Key

One can infer from Figure 3 that most of the data chose has songs written in the F major scale.

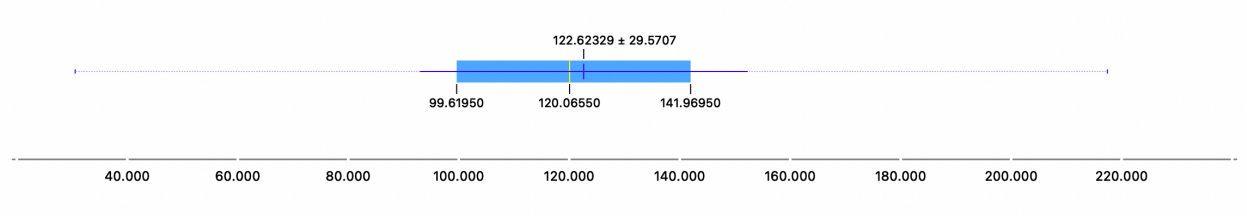

Figure 4: Tempo

From the Box Plot in Figure 4, one can see that most of the music data revolves around fast paced songs.

Figures 5 - 13 show histograms of all the other attributes:

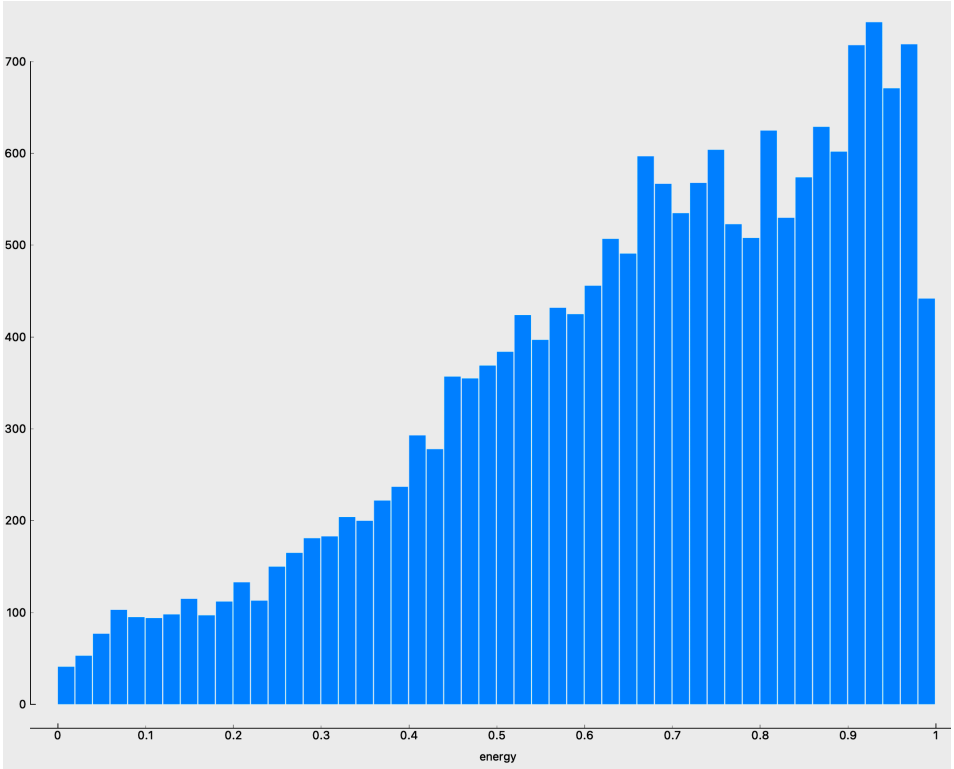

Figure 5. Energy

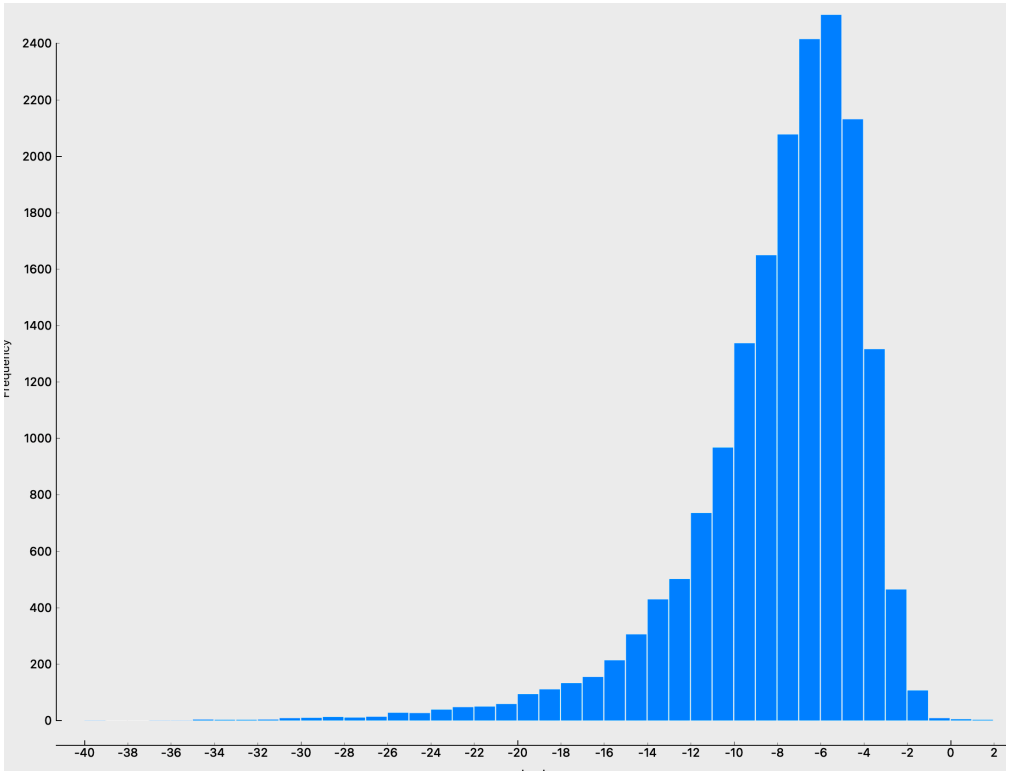

Fig 6: Loudness

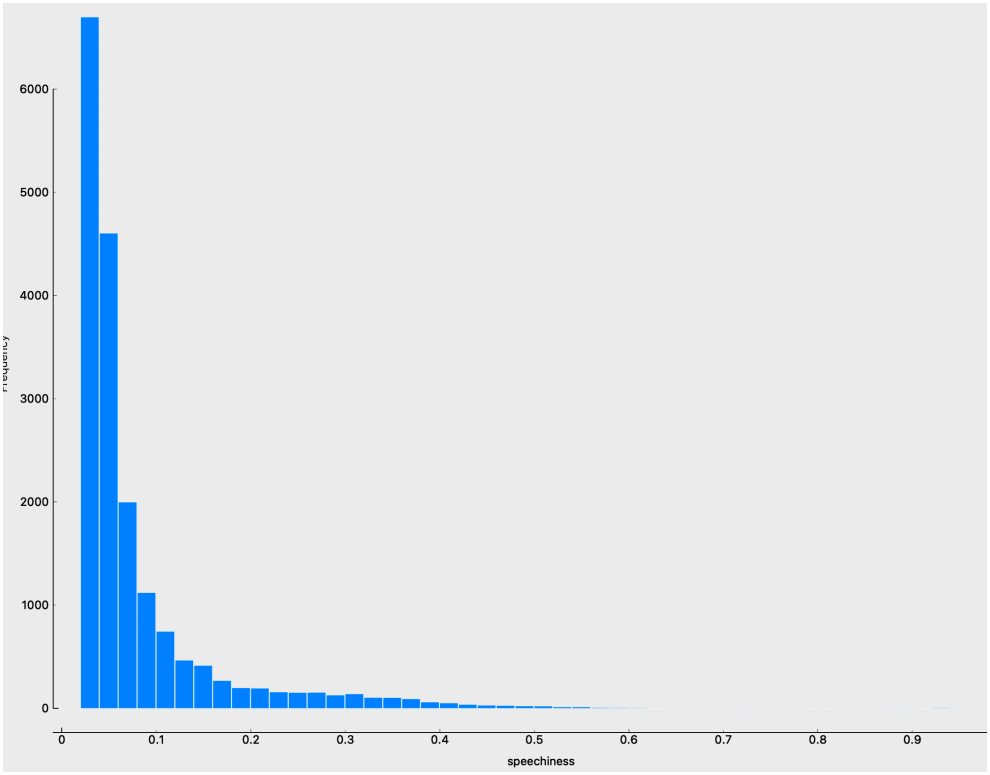

Fig 7: Speechiness

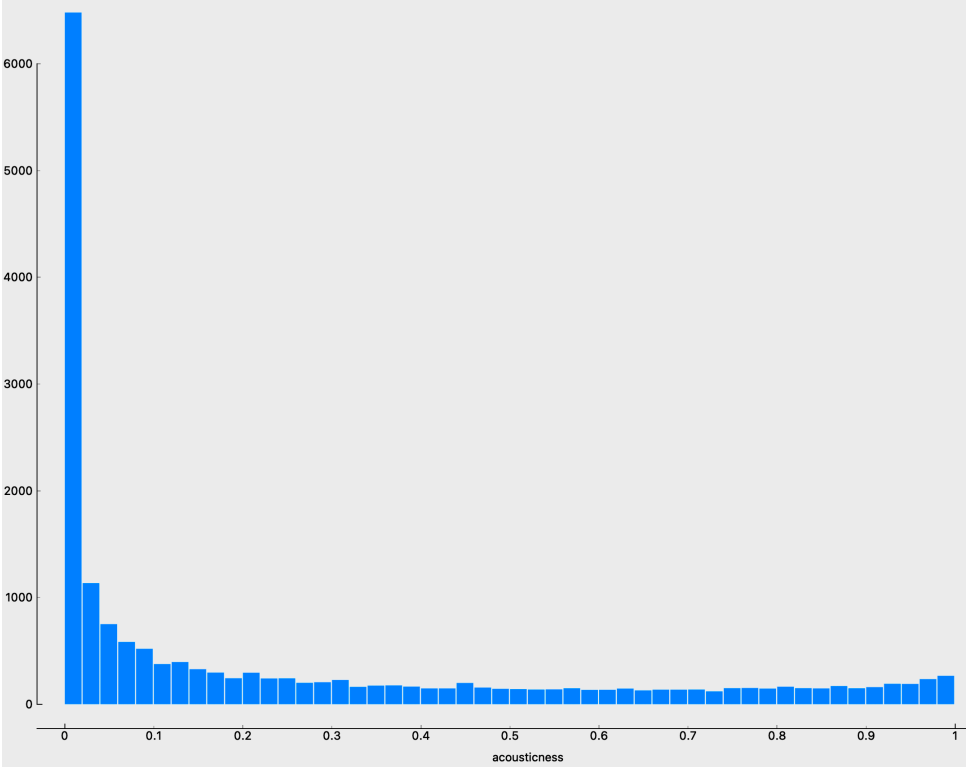

Fig 9: Acousticness

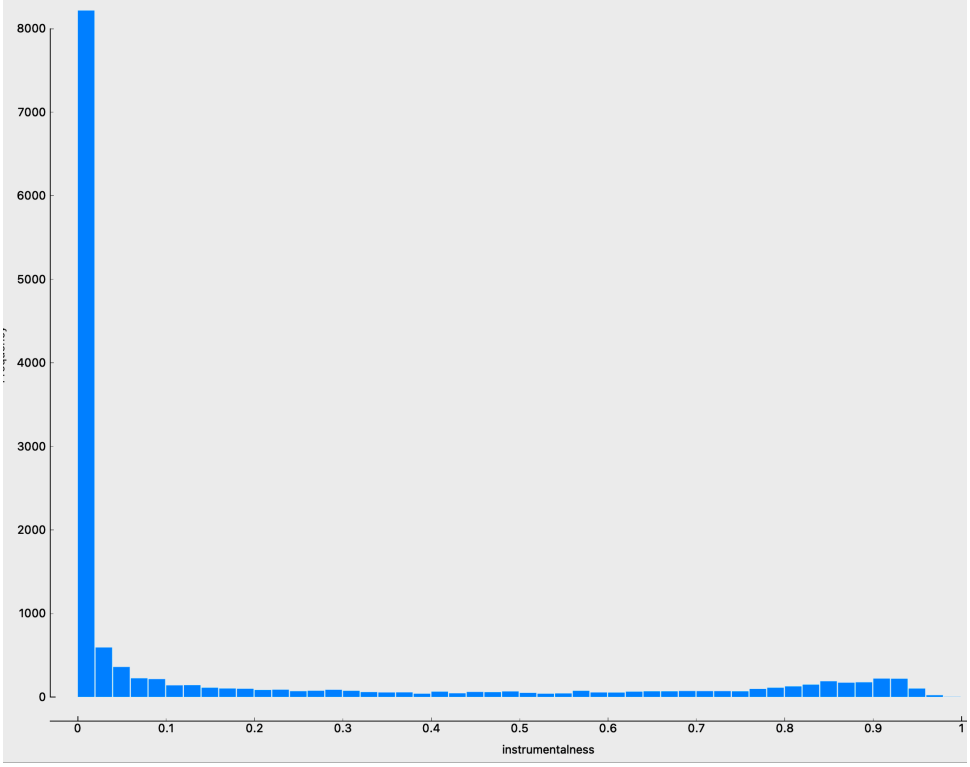

Fig 10: Liveness

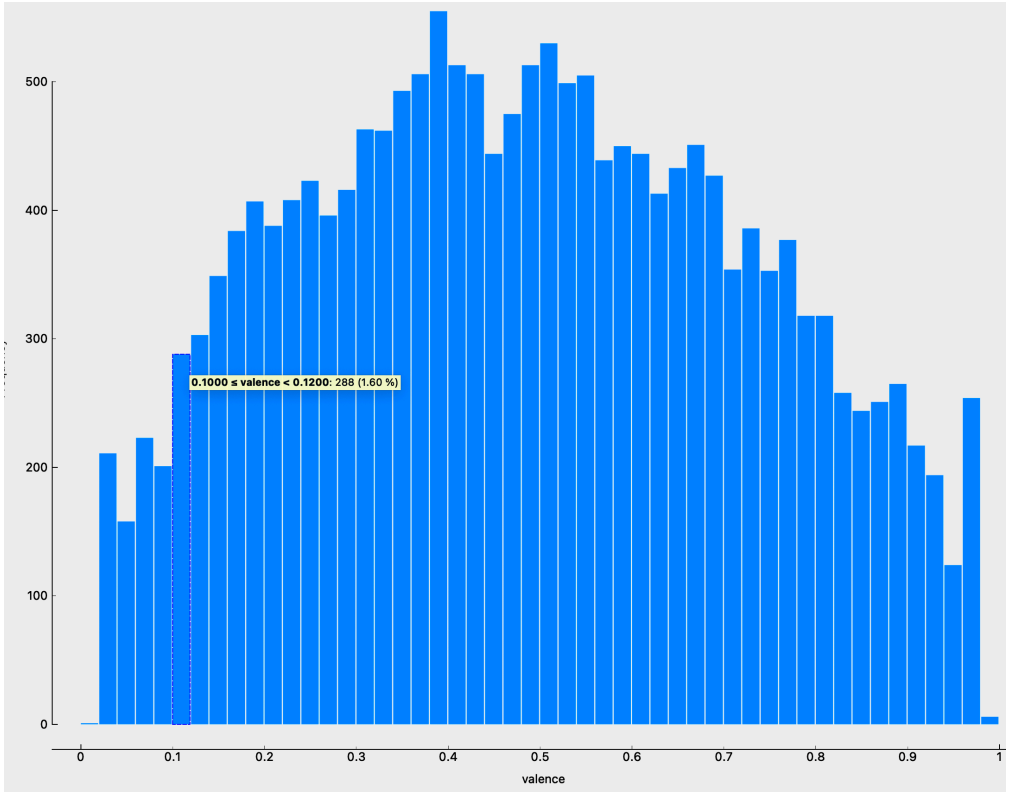

Fig 11: Valence

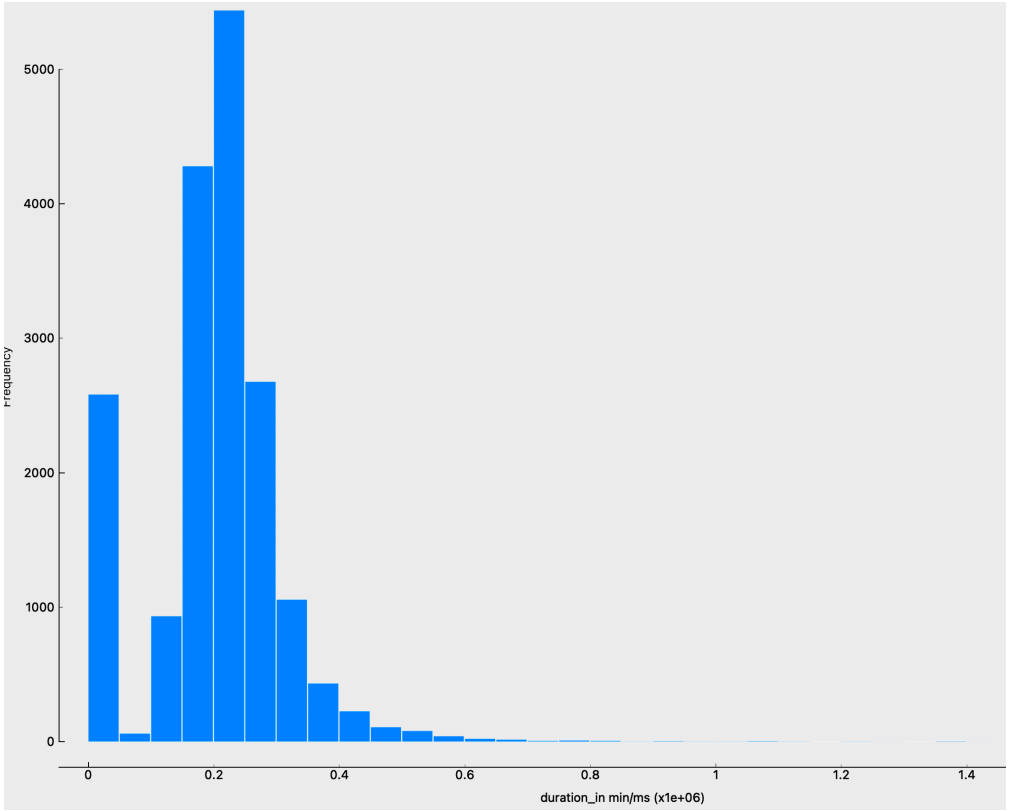

Fig 12: duration

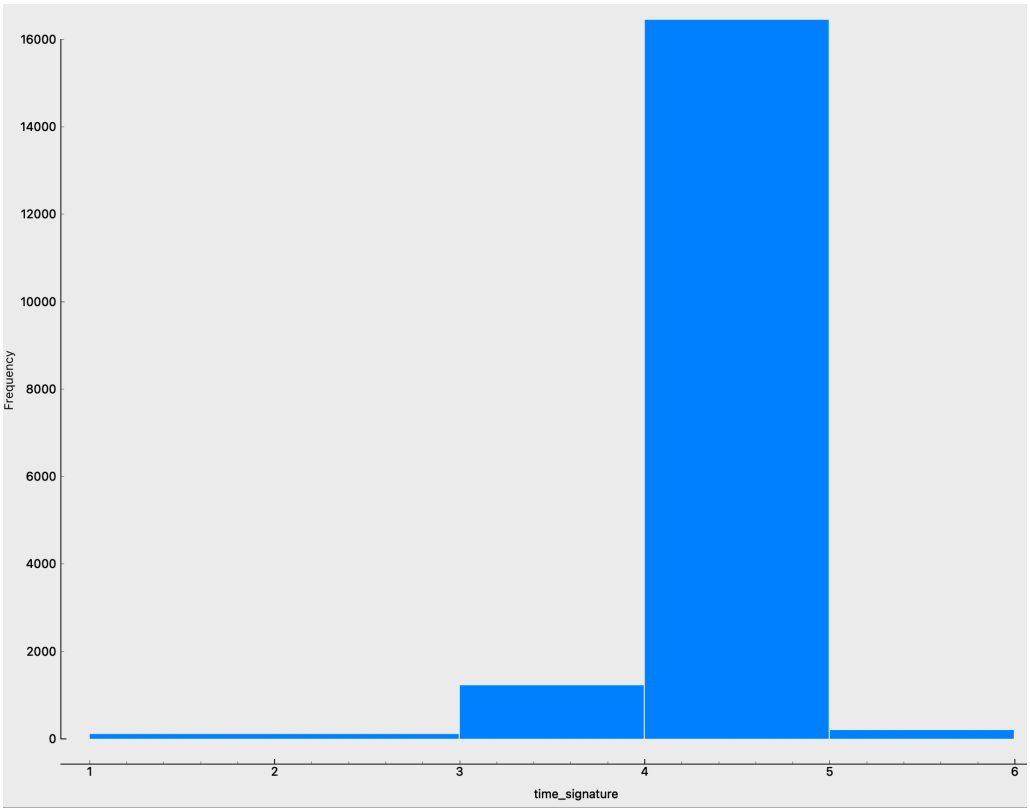

Figure 13: Time Signature

Methodology and Approach

My aim is to use density based clustering to find these outlier and then attempt to re-classify these outliers. I also do plan on extending this project further and rules of association between the attributes and genres.