Lab 5 – MATH 240 – Computational Statistics

Caroline Devine
Colgate University
Math Department
cdevine@colgate.edu

2/25/2025

Abstract

This lab is an extension of Lab 2 which focused on how to automate data extraction for .WAV files through batch processing as well as extracting and analyzing musical features from `.JSON` data. The extension is outlined in Task 3 under the methods section which compiles data from multiple tracks using the Essentia model and LIWC text analysis. The goal is to use the methods below to create data frames that can be used to facilitate analysis of musical influence.

Keywords: Data Collection, Lists, Batch Files, For Loops, JSON package

1 Introduction

The three bands: The Front Bottoms, Manchester Orchestra, and All Get Out collaborated on a song called "Allentown"(Ross, 2018) in 2018. This project aims at answering the research question of which band contributed most to this song. To analytically determine this, we will analyze 180 tracks and "Allentown" to examine the lyrical and sound features. The goal is to learn how to install, load, and learn how to use libraries; work with character objects; code `for()` loops; and access elements of vectors and lists.

2 Methods

2.1 Task 1: Building a Batch File for Data Processing

We analyzed a folder called "Music" containing songs from two artists, OfficeStuff and PeopleStuff, using `R`. We utilized the `list.dirs()` function to find album folders, and the `stringr` package(Wickham, 2023) allowed us to count and extract details from the .WAV files. We converted these details into .JSON format and saved the commands in a text file (`batfile.txt`). A `for()` loop automated the process, making it faster and easier to repeat in the future.

2.2 Task 2: Compiling Data from Essentia

We extracted key musical features from the song "Au Revoir (Adios)" using its .JSON file. Expanding this process, we iterated over 181 .JSON files from the Essentia Extractor Data, which conducts spectrogram analysis on music tracks. The extracted features were compiled into a dataframe alongside artist, album, and track metadata for further analysis.

2.3 Task 3: Load and Clean Data from Essentia Models

The dataset from the Essentia Model Data was processed to extract relevant musical features and metadata. Valence and arousal values were averaged across multiple sets. Key musical attributes were computed and refined, including renaming variables for clarity. Irrelevant columns were removed, resulting in a structured dataset optimized for further analysis. We used

The Essentia Model Data was obtained through using Python to interact with Essentia Models to collect data about what the music sounds like in more human terms. Using the provided .csv file, `EssentiaModelOutput.csv`, we used three datasets: DEAM, emoMusic, and MuSe to average each of their valance and arousal. New columns were created using `rowMeans()` of key musical features using different extractors such as Discogs-EffNet and MSD-MusiCNN in a similar manner as the valence and arousal columns. We renamed a column for clarity purposes changing `eff_timbre_bright` to `timbreBright`. Lastly, we isolated the desired columns which included the features created, renamed, and the artist, album, and track columns. This removed the unnecessary columns for future efficient analysis, leaving a clean data frame.

2.3.1 LIWC Text Analysis Tool

To analyze the lyrics of the tracks, we utilized a text analysis tool called LIWC which provides features that describes thoughts, feelings, and personality traits based on the language used. We loaded the `LIWCOutput.csv`. We merged the data from Task 1, the Essentia models, and the LIWC into one data frame using `merge()` function. The three data frames merged are called `df, cleaned.essentia.model, and LIWCOutputdf`. The resulting data frame consisted of 181 rows and 140 columns ensuring no column duplication or omission. Additionally, we renamed `function.` to `funct` because using `function` as a column name can result in issues in coding within `R`.

Lastly, we wrote two .csv files with one containing all tracks except "Allentown" called `trainingdata.csv` and the other containing only "Allentown" called `testingdata.csv`.

This is useful to evaluate the information solely based on "Allentown" as the initial research question calls for.

2.4 Task 4: Summarizing the Data

Importing a provided extended dataset containing 67 features from Essentia's music extractor, 14 features from Essentia models, 118 features from LIWC, and two additional variables from the bing sentiment lexicon(Hu and Liu, 2004) into `R` . This task implements `tidyverse`(Wickham et al., 2019) syntax, a summary of data numerically and visually, and a potential answer to the research question: which artist had a bigger impact on the track "Allentown"(Ross, 2018)?

Two datasets imported were read from .csv files, the main dataset and a subset specific to Allentown. We created a function, `range.allentown`, to compute summary statistics calculating the minimum, first quartile (Q1), third quartile (Q3), interquartile range (IQR), as well as lower (LF) and upper fences (UF) based on 1.5 times the IQR for a given feature grouped by artist. For each artist, the function calculates the Allentown-specific value for the feature which is then compared against these statistics to classify it as "Out of Range," "Outlying," or "Within Range." The resulting summary table, which includes the artist name and statistics, is used to analyze differences between artists.

We then identified numeric feature from the main dataset using R's `is.numeric` function with `sapply`, storing the names of the columns in a vector. At the same time, the function `range.allentown` was applied to each numeric feature using `lapply` to compute summary statistics and range classifications for the Allentown subset. All the outputs were combined into one summary table using `bind_rows` for further analysis.

To determine which features were most influential, we initially examined 20 candidate features across three bands, noting that only one band out of the three consistently exhibited values classified as "Within Range." Based on this observation, we selected eight features: dissonance, average loudness, chords strength, spectral rolloff, perception, positive words, OtherP, and conjunctions. We created two tables: one (`isolated.feature.sum`) displaying detailed statistics (minimum, lower fence, upper fence, and, maximum) for each feature per artist (see Appendix), and another providing a concise view of the artist, feature, and description. Both tables were formatted into LaTeX using the `xtable` package(Dahl et al., 2019).

artist	feature	description
All Get Out	spectral_rolloff	Out of Range
Manchester Orchestra	spectral_rolloff	Within Range
The Front Bottoms	spectral_rolloff	Out of Range
All Get Out	dissonance	Outlying
Manchester Orchestra	dissonance	Within Range
The Front Bottoms	dissonance	Out of Range
All Get Out	average_loudness	Outlying
Manchester Orchestra	average_loudness	Within Range
The Front Bottoms	average_loudness	Outlying
All Get Out	chords_strength	Outlying
Manchester Orchestra	chords_strength	Within Range
The Front Bottoms	chords_strength	Out of Range

Table 1: Sound Features

Lastly, the eight features selected, four are from Essentia which represents sound information and four are from LIWC which is representative of lyrical information. We created Violin and box plots were created with `ggplot2`(Wickham, 2016). For sound, plots were generated for dissonance, average loudness, chords strength, and spectral rolloff; for lyrics, similar plots were made for perception, positive words, OtherP, and conjunctions, with each plot including a horizontal line indicating the corresponding "Allentown" value. We combined the two plots using the `patchwork` package(Pedersen, 2024) to produce visualizations for sound and lyrical features.

3 Results

For Task 1, we successfully analyzed .WAV files, storing the results in .JSON files. For Task 2 and 3, we extracted key musical features from 181 tracks,cleaned and organized the data from Essentia model's .JSON outputs, combined those data points with LIWC text analysis, and created training and testing data sets(Bogdanov et al., 2013). Task 4 showed that Manchester Orchestra predominantly influenced sound features, while The Foot Bottoms had the greatest impact on lyrical features.

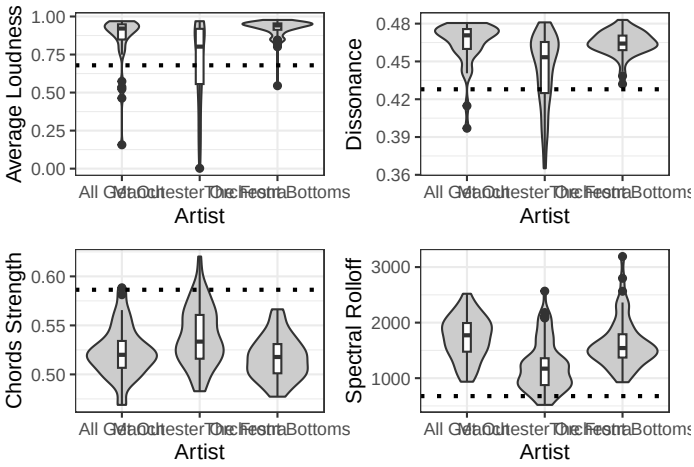

Figure 1: Sound Features

Figure 1 shows the four sound features with the black dotted line intersecting with the Manchester Orchestra in all four. This is replicated for the lyrical features in Figure 2 in the appendix.

4 Discussion

This process is now automated and can be replicated with other files as well. This creates an efficient way to compare larger data sets of multiple artists or albums. The integrated, structured data frame allows for specific analysis to address research question in future work.

References

Bogdanov, D., Wack, N., Gómez Gutiérrez, E., Gulati, S., Boyer, H., Mayor, O., Roma Trepat, G., Salamon, J., Zapata González, J. R., Serra, X., et al. (2013). Essentia: An audio analysis library for music information retrieval. pages 493–498.

Dahl, D. B., Scott, D., Roosen, C., Magnusson, A., and Swinton, J. (2019). *xtable: Export Tables to LaTeX or HTML*. R package version 1.8-4.

Hu, M. and Liu, B. (2004). Mining and summarizing customer reviews. pages 168–177.

Pedersen, T. L. (2024). *patchwork: The Composer of Plots*. R package version 1.3.0.

Ross, A. R. (2018). Manchester orchestra and the front bottoms are finally together on "allentown". *Vice*.

Wickham, H. (2016). *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York.

Wickham, H. (2023). *stringr: Simple, Consistent Wrappers for Common String Operations*. R package version 1.5.1.

Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L. D., François, R., Grolemund, G., Hayes, A., Henry, L., Hester, J., Kuhn, M., Pedersen, T. L., Miller, E., Bache, S. M., Müller, K., Ooms, J., Robinson, D., Seidel, D. P., Spinu, V., Takahashi, K., Vaughan, D., Wilke, C., Woo, K., and Yutani, H. (2019). Welcome to the tidyverse. *Journal of Open Source Software*, 4(43):1686.

5 Appendix

artist	feature	description
All Get Out	conj	Out of Range
Manchester Orchestra	conj	Outlying
The Front Bottoms	conj	Within Range
All Get Out	Perception	Out of Range
Manchester Orchestra	Perception	Within Range
The Front Bottoms	Perception	Out of Range
All Get Out	OtherP	Outlying
Manchester Orchestra	OtherP	Outlying
The Front Bottoms	OtherP	Within Range
All Get Out	positivewords	Outlying
Manchester Orchestra	positivewords	Outlying
The Front Bottoms	positivewords	Within Range

Table 2: Lyrical Features

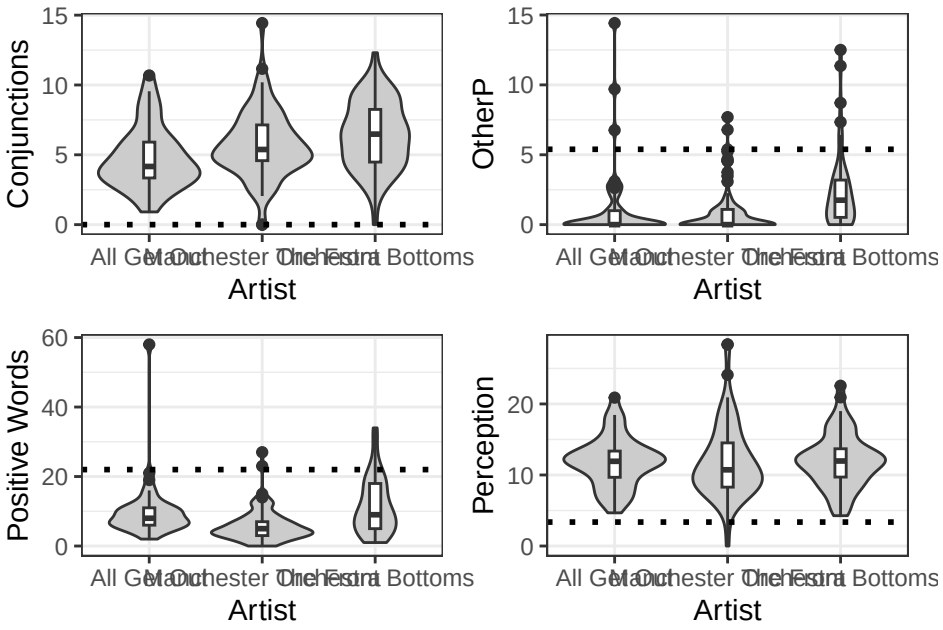

Figure 2: Lyrical Features

Table 3: Summary Table of 8 selected features

artist	min	LF	UF	max	description	feature
All Get Out	935.91	701.91	2767.30	2520.04	Out of Range	spectral_rolloff
Manchester Orchestra	518.87	151.27	2083.17	2566.67	Within Range	spectral_rolloff
The Front Bottoms	927.04	740.58	2421.46	3190.29	Out of Range	spectral_rolloff
All Get Out	0.40	0.44	0.50	0.48	Outlying	dissonance
Manchester Orchestra	0.37	0.36	0.53	0.48	Within Range	dissonance
The Front Bottoms	0.43	0.44	0.49	0.48	Out of Range	dissonance
All Get Out	0.16	0.70	1.10	0.97	Outlying	average_loudness
Manchester Orchestra	0.00	0.01	1.46	0.97	Within Range	average_loudness
The Front Bottoms	0.55	0.85	1.02	0.98	Outlying	average_loudness
All Get Out	0.47	0.47	0.58	0.59	Outlying	chords_strength
Manchester Orchestra	0.48	0.45	0.63	0.62	Within Range	chords_strength
The Front Bottoms	0.48	0.46	0.58	0.57	Out of Range	chords_strength
All Get Out	0.91	-0.51	9.76	10.68	Out of Range	conj
Manchester Orchestra	0.00	0.74	10.98	14.43	Outlying	conj
The Front Bottoms	0.00	-1.17	13.90	12.31	Within Range	conj
All Get Out	4.67	4.14	18.91	20.89	Out of Range	Perception
Manchester Orchestra	0.00	-1.06	23.87	28.37	Within Range	Perception
The Front Bottoms	4.27	3.74	19.66	22.56	Out of Range	Perception
All Get Out	0.00	-1.50	2.50	14.42	Outlying	OtherP
Manchester Orchestra	0.00	-1.64	2.73	7.69	Outlying	OtherP
The Front Bottoms	0.00	-3.48	7.20	12.50	Within Range	OtherP
All Get Out	2.00	-1.50	18.50	58.00	Outlying	positivewords
Manchester Orchestra	0.00	-3.00	13.00	27.00	Outlying	positivewords
The Front Bottoms	1.00	-14.50	37.50	34.00	Within Range	positivewords