Lab 5 – MATH 240 – Computational Statistics

Caroline Devine
Colgate University
Math Department
cdevine@colgate.edu

2/25/2025

Abstract

This project produces data extraction by processing .WAV files to `.JSON` data and analyzing musical features via batch files and loops. We used Essentia(Bogdanov et al., 2013) for sound analysis and LIWC(Boyd et al., 2022) for lyrical examination, compiling data from 181 tracks were compiled into one dataset. We completed 5 tasks involving extracting, cleaning, merging, and visualizing key features to assess artist influence on the track "Allentown." Our results indicate that Manchester Orchestra predominantly influenced the sound, while The Front Bottoms influenced the lyrics.

Keywords: Data Collection & Extraction, Batch Processing, Data Cleaning, Violin Plots

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" through the lyrical and sound features.

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 `jsonlite` package used for reading and parsing .JSON files in R (Ooms, 2014). Expanding this process, we iterated over 181 .JSON files from the Essentia Extractor Data, which conducts spectrogram analysis on music tracks(Bogdanov et al., 2013). The extracted features were compiled into a dataframe with artist, album, and track data for further analysis.

2.3 Task 3: Load and Clean Data from Essentia Models

We used the Essentia Model Data which provides data about what the music sounds like in more human terms. Valence and arousal values were averaged across multiple sets. Key musical attributes were created using different extractors, including renaming features for clarity. Also, irrelevant columns were removed, resulting in a a clean data frame.

2.4 Task 4: LIWC Text Analysis Tool

To analyze track lyrics, we used the LIWC text analysis tool, which extracts linguistic features related to thoughts, feelings, and personality traits(Boyd et al., 2022). The processed data was then loaded for analysis. Data from Task 1, the Essentia models, and LIWC were merged into a single data frame, ensuring no duplication or omission. The final dataset contained 181 rows and 140 columns. To prevent coding conflicts in R, the column `function.` was renamed to `funct`.

2.5 Task 5: Summarizing the Data

We imported 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)?

For each numerical feature, summary statistics (minimum, first quartile, third quartile, interquartile range, and fences) were computed, grouped by artist.Each Allentown-specific value was then classified as "Out of Range", "Outlying", or "Within Range", allowing for a comparison of artist differences.

To identify the most influential features, we initially evaluated 20 features, noting only one band consistently exhibited

"Within Range" values, and selected eight (dissonance, average loudness, chords strength, spectral rolloff, perception, positive words, OtherP, and conjunctions). Detailed statistics for these features were formatted into Table 3 (see Appendix) and Table 1 (see below) provided a concise view of the artist, feature, and description 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

We selected eight features were selected—four from Essentia (sound) and four from LIWC (lyrics). Violin and box plots for sound (dissonance, average loudness, chords strength, spectral rolloff) and lyrics (perception, positive words, OtherP, conjunctions) were created using `ggplot2` (Wickham, 2016), with each plot including a horizontal line for the corresponding "Allentown" value. We combined the plots using the `patchwork` package (Pedersen, 2024) to visualize both 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)(Boyd et al., 2022). Task 4 showed that Manchester Orchestra predominantly influenced sound features, while The Foot Bottoms had the greatest impact on lyrical features.

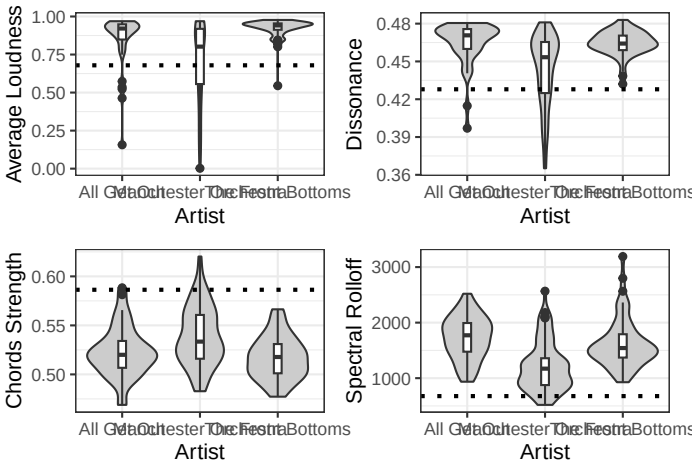

Figure 1: Sound Features

The violin plots shows that, for each of the four features, "Allentown's" value (indicated by the dashed line) aligns most closely with the distribution of Manchester Orchestra, suggesting that the track's sounds are most similar to this band out of the three artists. A similar plot is replicated for the lyrical features in Figure 2 in the appendix. The lyrical feature plot indicates that The Front Bottoms has the greatest influence on the lyrical aspects of "Allentown", particularly in conjunctions, positive words, and other parts of speech, despite the dashed line not aligning perfectly with the center of the distribution. Perception, however, is an outlier, where Manchester Orchestra's influence is most prominent.

4 Discussion

With the eight identified features, we successfully identified the influence of a particular band on the track "Allentown", with Manchester Orchestra influencing the sound and The Front Bottoms influencing the lyrics. This can be replicated to see how all 20 of the original selected features compare to these results. Manually, we identified that Manchester Orchestra influenced all the 12 other features selected. We did not select the features which two bands exhibited "Within Range" values (21 total) which could also potentially shed light on more accurate results. We chose to look at the "Within Range" values, but potential other avenues to dissect which artist had the most influence could be focused on different summary variables such as mean, average, standard deviation. This project can be replicated by others with only slight manipulations to the code in `R` for other musical analysis with different artists and tracks.

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.

Boyd, R. L., Ashokkumar, A., Seraj, S., and Pennebaker, J. W. (2022). *The development and psychometric properties of LIWC-22.* Austin, TX: University of Texas at Austin, 10.

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.

Ooms, J. (2014). The jsonlite package: A practical and consistent mapping between json data and r objects. *arXiv:1403.2805 [stat.CO].*

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

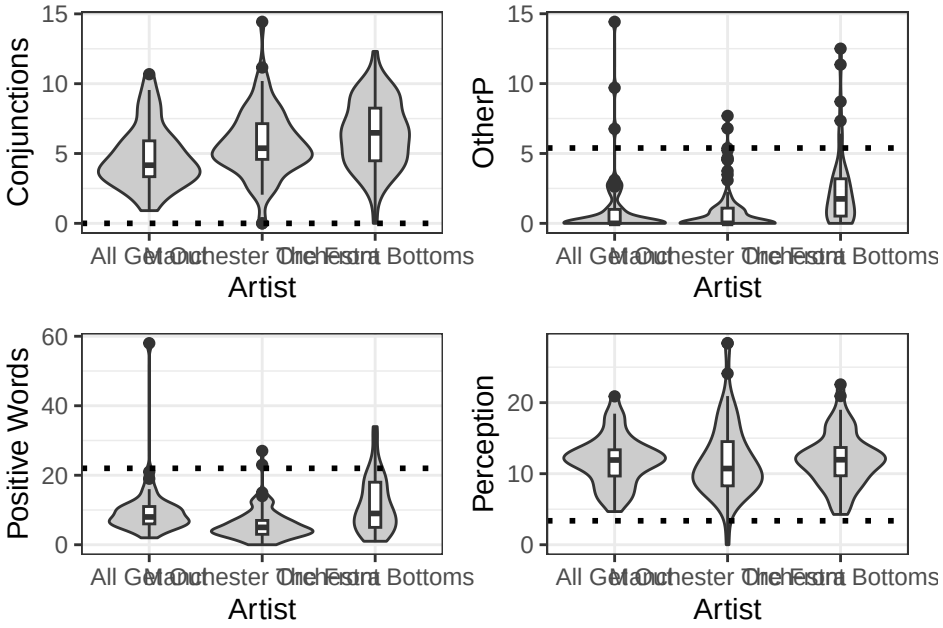

Figure 2: Lyrical Features

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