# Lab 05 – MATH 240 – Computational Statistics

Reagan Sernick
Affiliation
Department
rsernick@colgate.edu

2/27/25

**Abstract**

In this lab, we took the data we extracted from last lab and interpreted it in terms of wether the data for *Allentown* was within range, outlying, or out of range of the data from each of the artists that contributed. In the end this will give us an idea of which artist contributed the most.

**Keywords:** `tidyverse`, `ggplot2`

## 1 Introduction

The purpose of this lab, and the two leading up to it, was to determine which artist contributed the most on *Allentown* of the contributing artists. To do this, we had to take the data from the previous lab and interpret it in a meaningful way.

## 2 Methods

The data we used from this lab came from Essentia (Bogdanov et al., 2013), Essentia Models (Alonso-Jiménez et al., 2020), and Linguistic Inquiry and Word Count (LIWC) (Boyd et al., 2022). These programs help model data from a song's waveform and lyrics in a quantitative way. To answer our research question, we needed to determine whether the value for a specific feature output by the models was within range, outliyng, or out of the range of the values from each artist.

### 2.1 Summarize Data for Features by Artist

In order to determine if *Allentown* is within range, outlying, or out of range of every feature I needed to create a function that takes the feature name as a parameter. First in the function I needed to calculate the IQR, minimum, lower fence, upper fence, and maximum of the data for each artist. This was made easier using the `group_by()` and `summarize()` functions in `tidyverse` (Wickham et al., 2019).

Now with these values I calculated whether the feature value for *Allentown* was out of range (<min or >max), outlying (<LF or >UF), or within range, and stored values for each feature as such.

### 2.2 Counts

Now with the range data for each feature I needed to calculate the number of times within range, outlying, and out of range appeared in an artist's row. By using a `for()` loop that ran for every feature, `case_when()`, and `rowwise()` (another tidyverse function) I was able to get a count of every feature's range data by artist.

### 2.3 Separating Data

To help answer the question, I wanted to see how the data from the song's wave forms differed from the data from the song's lyrics. Since I knew that the Essentia data was from the song's wave forms I repeated the counting step for the first 78 features (all of the features from Essentia/Essentia Models) labeling it Musical Data, and repeated the counting step for every feature after the 78th (all features from LIWC) labeling it lyrical data.

## 3 Results

Using `ggplot2`, (Wickham, 2016) I tested creating multiple column plots, but ultimately ended up dividing the data into All Data, Musical Data, and Lyrical Data and plotting artist against the within range count. I thought this was the best way to display the data and answer the question because it gives two different answers. In addition to the graphs, I used xtable (Dahl et al., 2019) to create tables of the range data for All Data, Musical Data, and Lyrical Data.

**All Data:** As shown in All Data graph and All Data table, *Allentown* has the most features within range of Manchester Orchestra.

**Musical Data:** As shown in Musical Data graph and Musical Data table, *Allentown* has the most musical features within range of Manchester Orchestra.

**Lyrical Data:** As shown in Lyrical Data graph and Lyrical Data table, *Allentown* has the most lyrical features within range of The Front Bottoms

# 4 Discussion

**All Data:** The results from the analysis of all features suggest that *Allentown* is most similar to Manchester Orchestra, with 183 features within range. This indicates that the overall characteristics of *Allentown*, considering both musical and lyrical elements, align most with Manchester Orchestra compared to the other artists. The Front Bottoms and All Get Out exhibit fewer similarities, as reflected in their lower counts of within-range features and higher instances of outlying or out-of-range values.

**Musical Data:** The analysis of musical features reinforces the results of the previous analysis, with 76 features within range of Manchester Orchestra. It also shows that *Allentown* had the least amount of features within range of The Front Bottoms (47) and the most out of range (26). From this it is reasonable to conclude that *Allentown* has a musical style most similar to Manchester Orchestra and least similar to The Front Bottoms.

**Lyrical Data:** Interestingly, the lyrical feature analysis reveals that The Front Bottoms have the highest number of within range features (109). This suggests that, while *Allentown* aligns musically with Manchester Orchestra, its lyrical characteristics are more similar to The Front Bottoms. This distinction indicates that different elements of the song may be influenced by different artists, with Manchester Orchestra shaping its musical aspects and The Front Bottoms influencing its lyrical composition.

# References


Alonso-Jiménez, P., Bogdanov, D., Pons, J., and Serra, X. (2020). Tensorflow audio models in essentia. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 266–270. IEEE.

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. In Britto, A., Gouyon, F., and Dixon, S., editors, *14th Conference of the International Society for Music Information Retrieval (ISMIR)*, pages 493–498, Curitiba, Brazil. International Society for Music Information Retrieval (ISMIR).

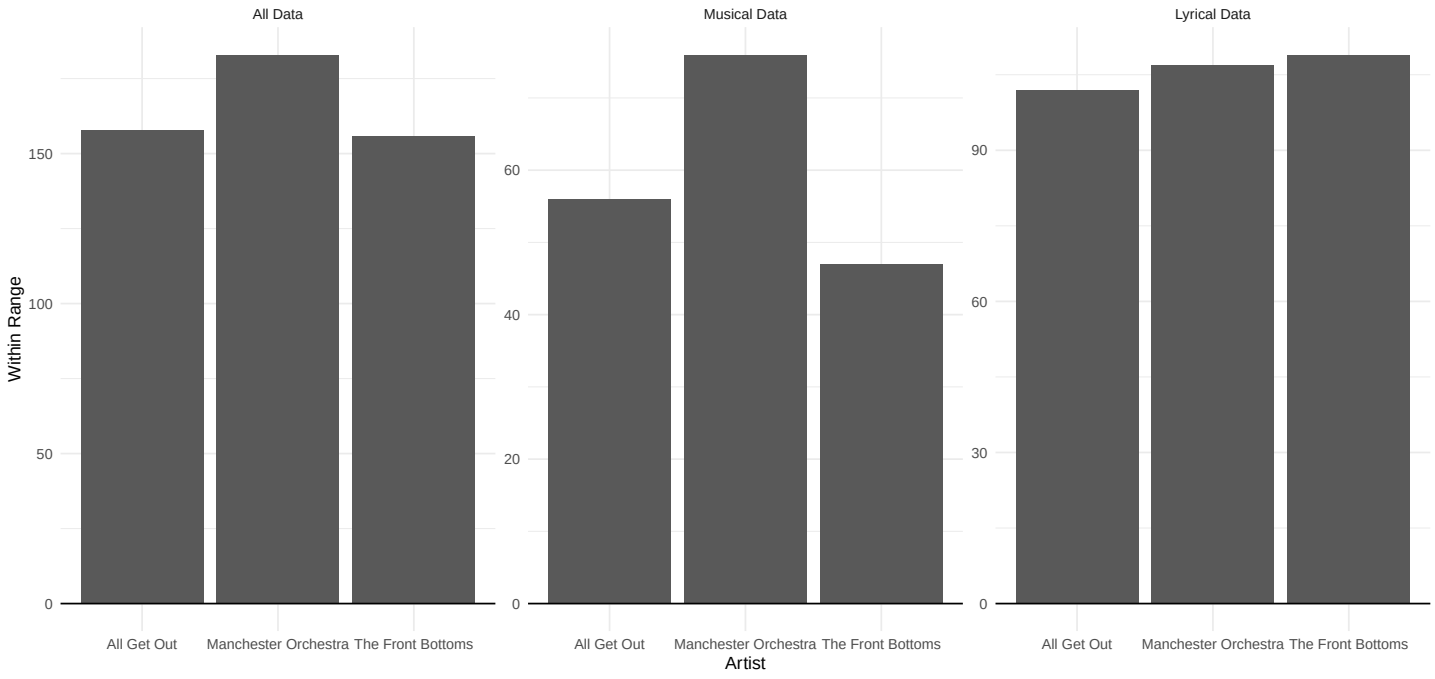
Boyd, R. L., Ashokkumar, A., Seraj, S., and Pennebaker, J. W. (2022). *The development and psychometric properties of LIWC-22*. Austin, TX.

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.

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

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.

# Appendix




Comparison of All Data, Musical Data, and Lyrical Data

**All Data**

| | Artist | Within Range | Outlying | Out of Range |
|---|---|---|---|---|
| 1 | All Get Out | 158 | 17 | 22 |
| 2 | Manchester Orchestra | 183 | 11 | 3 |
| 3 | The Front Bottoms | 156 | 11 | 30 |

Range of all Features Similar to *Allentown* by Artist

**Musical Data**

| | Artist | Within Range | Outlying | Out of Range |
|---|---|---|---|---|
| 1 | All Get Out | 56 | 10 | 12 |
| 2 | Manchester Orchestra | 76 | 2 | 0 |
| 3 | The Front Bottoms | 47 | 5 | 26 |

Range of Musical Features Similar to *Allentown* by Artist

**Lyrical Data**

| | Artist | Within Range | Outlying | Out of Range |
|---|---|---|---|---|
| 1 | All Get Out | 102 | 7 | 10 |
| 2 | Manchester Orchestra | 107 | 9 | 3 |
| 3 | The Front Bottoms | 109 | 6 | 4 |

Range of Lyrical Features Similar to *Allentown* by Artist