Lab 5 – MATH 240 – Computational Statistics

Camilo
Colgate University
Mathematics Department
cgranadacossio@colgate.edu

2/25/25

Abstract

This lab applied statistical analysis techniques using `tidyverse` (Wickham et al., 2019) in `R` to determine which band has the largest influence on the collaborative song *Allentown* by the All Get Out, The Front Bottoms, and Manchester Orchestra. Using datasets sets from `Essentia` (Bogdanov et al., 2013) and `LIWC` (Pennebaker et al., 2015), I extract musical features to classify the song's similarity to each band. The analysis utilizes **outlier detection, summary statistics, and data visualization techniques** to provide insights

Keywords: Statistical analysis; Data visualization; Outlier detection; `tidyverse`; Musical feature

1 Introduction

In 2018, *Allentown*, a collaborative song by The Front Bottoms, Manchester Orchestra, and All Get Out, was released. The goal of this lab is to determine which band contributed more significantly to the song's composition using statistical techniques. The lab focuses on extracting numerical features from `Essentia's` dataset, summarizing and comparing these features across the three bands, and using `ggplot2` (Wickham, 2016) for visualization. By implementing boxplots, scatter plots, and summary statistics, I identified patterns that highlight *Allentown's* alignment with each band's musical style.

2 Methods

The dataset includes key features such as `overall_loudness, tempo, danceability, and emotion`. The function `out()` was developed to generalize statistical assessements across features, computing summary statistics and classifying outliers. This was done using `group_by()` and `summarize()`. I computed minimum and maximum values per artist and determined the lower and upper fences ($Q\_1 - 1.5 \times IQR, Q\_3 + 1.5 \times IQR$) to detect outliers. I applied `mutate()` to indicate whether *Allentown* was **out of range**, an **outlier**, or **within range**. A filtered DataFrame stored the statistical results for all the numerical features. Visualization of the results were created through box plots comparing *Allentown's* feature values to each band's distribution and scatter plots highlighting *Allentown's* placement relative to the bands. The `facet_wrap()` function was used to arrange the plots together.

3 Results

Analysis of teh dataset reveals that *Allentown's* `overall_loudness, tempo, danceability, and emotion` align more closely with Manchester Orchesta's style.

4 Discussion

In the future, plots with more variables could be made to better visualize the correlations between artists differing audio features.

References

Bogdanov, D., Wack, N., Gómez, E., Gulati, S., Boyer, H., Mayor, O., et al. (2013). Essentia: an audio analysis library for music information retrieval. In *Proceedings of the 14th International Society for Music Information Retrieval Conference (ISMIR)*, pages 493–498.

Pennebaker, J. W., Booth, R. J., Boyd, R. L., and Francis, M. E. (2015). *Linguistic Inquiry and Word Count: LIWC2015*. Pennebaker Conglomerates, Austin, TX.

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.

5 Appendix