Lab 05 – MATH 240 – Computational Statistics

Danny Molyneux
Colgate University
Mathematics
dmolyneux@colgate.edu

Feb 10 2024

Abstract

In this lab we are tasked with using tidyverse (Wickham et al., 2019) to gather stats for each musical feature for all three artists (The Front Bottoms, Manchester Orchestra, and All Get Out). We also extract these feature values for their combined song, Allentown. Then, use this information to answer the overarching question: Which band contributed most to the song Allentown?

Keywords: tidyverse; summarizing; analyzing; plotting

1 Introduction

The goal here is to address the question "Which band contributed most to the song Allentown?". We have over 200 different features, for 181 different songs. The idea is to extract information on each band by looking at their batch of songs, and compare that information to the features of the song Allentown. Hopefully, we will see significant differences, and that one band has clearly more similar traits to that of Allentown. I will be sure to make various plots/graphs to get various insights and eventually come to a conclusion.

2 Methods

The data we are working with here is two .csv files, one with all of the features for Allentown, and one for the features for the other 180 Essentia (Bogdanov and Serra, 2013) files. These features came from both Essentia and LIWC (Boyd and Pennebaker, 2022). Using (tidyverse), I use methods such as group_by() and summarize() to create a data frame containing a multitude of statistics for each feature, grouped by artist. So this will look at each artist, take all of their songs from the csv file, and summarize the features.

The mutate() function was very helpful as it allows me to create new columns that describe whether each artist is "Out of Range", "Outlying", or "Within Range" for each feature.

The xtable library (Dahl et al., 2019) allows me to create a table that summarizes select features, so that we can pick and choose which features we want to use to answer the overarching question.

Lastly, using ggplot (Wickham, 2016) is a way to really visualize our findings. Similarly to the xtable, we can plot specific features to see how the artists compare to Allentown's feature. Alternatively, we can use ggplot to compare the artists overall comparison's to Allentown, by calculating the proportion of each description (Out of Range, Outlying, Within Range) for each artist. I do this by creating a new data frame that for each artist, has both a count and proportion for each description type. We would plot this by putting the plot for each artist side-by-side.

3 Results

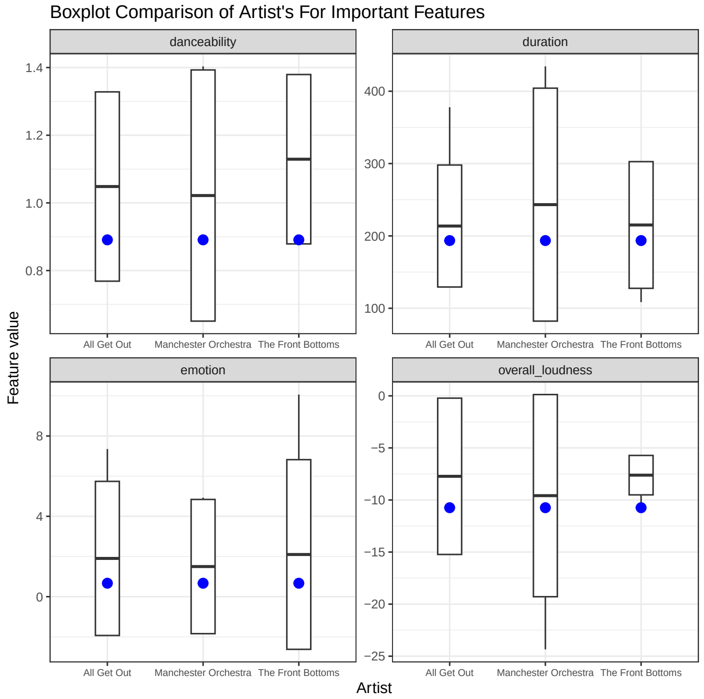
After collecting all of the data into my various data frames, here is the table I make for the four features that I especially wanted to analyze:

	artist	feature	description
1	All Get Out	overall_loudness	Within Range
2	Manchester Orchestra	overall_loudness	Within Range
3	The Front Bottoms	overall_loudness	Outlying
4	All Get Out	danceability	Within Range
5	Manchester Orchestra	danceability	Within Range
6	The Front Bottoms	danceability	Out of Range
7	All Get Out	duration	Within Range
8	Manchester Orchestra	duration	Within Range
9	The Front Bottoms	duration	Within Range
10	All Get Out	emotion	Within Range
11	Manchester Orchestra	emotion	Within Range
12	The Front Bottoms	emotion	Within Range

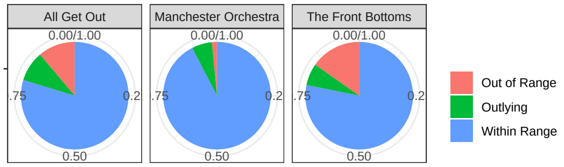
Table 1: Results of each essentia feature vs Allentown

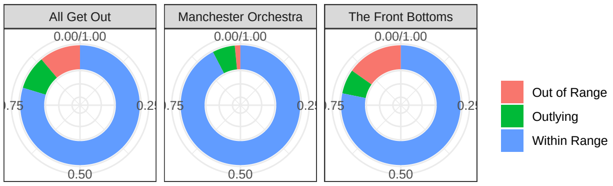
So although I liked the idea of using emotion and duration, they ended up not being very helpful, because each artist is within range for those features, meaning the Allentown feature is within their lower fence ($Q1 - 1.5IQR$) and upper fence ($Q3 + 1.5IQR$). So these two features don't seem to give us much insight at all. However, if you look at the box plot below, you will see that Allentown's duration is quite a bit further away from Manchester Orchestra's median than the others. This still doesn't tell us much, because it's one feature and they are all still relatively close to the median.

On the other hand, we learn a lot from the table and box plot about overall loudness and danceability. Allentown is Out of Range and Outlying in regards to The Front Bottoms' danceability and overall loudness, respectively. For the other two bands, both features are within range. So if you find these two traits valuable, it may lead you to believe that The Front Bottoms weren't as contributive to Allentown as Manchester Orchestra and All Get Out.

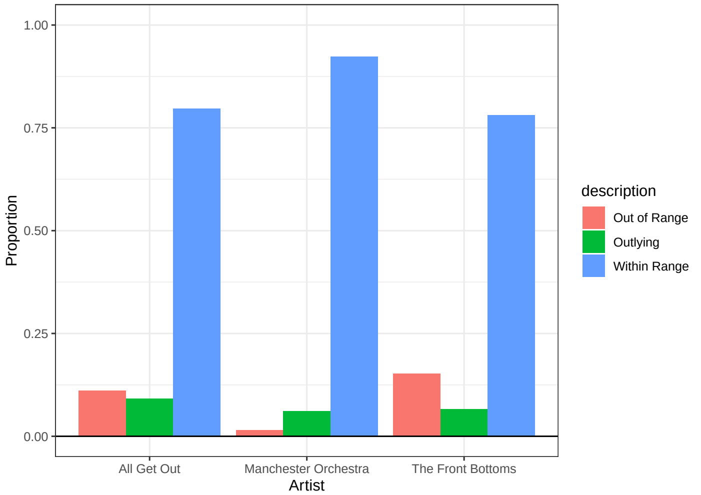


As I mentioned in the Methods section, I also use ggplot to analyze the artists as a whole, rather than for specific features. To do so, I made a few different side-by-side plots, comparing the proportions of descriptions for each artist. Let's start by looking at the pie chart and doughnut plot I have made, because they are quite similar graph types.





I really like these plots (doughnut more than pie) because they tell you in general, which bands were more often in the same range as Allentown, as well as the contrary. I think these plots are very much so in favor of Manchester Orchestra being the band that contributed most to Allentown. They have the highest proportion of "Within Range" and the lowest proportion of "Out of Range". In regards to the other two bands, it isn't super obvious, but All Get Out has a slightly higher proportion of 'Within Range" and definitely has a smaller proportion of "Out of Range". Also note that I am using proportions rather than just counting, because the sample size for each band is different. Let's look at one more plot in case the doughnut/pie plots were too difficult to visualize. Here is a column plot of the same data:



This tells the same story, but maybe a little more clearly.

4 Discussion

So those last three graphs definitely gave us some strong leads to answer this question, but is it conclusive? I'd say no. The reason is, we don't know exactly which features fall under which description for each artist. And how do we weigh each feature? I know I chose those four features earlier that I thought were especially important, but I am by no means a musical expert. For all we know, the four most significant features were within range for The Front Bottoms, and Out of Range for the other two bands. Is it likely? No. But it is possible.

The othr thing we must keep in mind is that even if Allentown is more similar to the sample of The Front Bottoms songs we have, doesn't mean they actually contributed more. If we kept increasing the sample size, we could probably say it with high confidence, but Allentown could always be an outlier. So how could we answer this question with more certainty other than increasing sample size?

References

Bogdanov, D. and Serra, X. (2013). Essentia: An audio analysis library for music information retrieval. pages 493–498.

Boyd, R. L. and Pennebaker, J. W. (2022). Liwc: The development and psychometric properties of liwc-22.

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.