Lab 05 – MATH 240 – Computational Statistics

Cristian Palmer
Student
Mathematics
cpalmer@colgate.edu

Abstract

For the past 3 weeks we have been working towards answering the question of which of three bands, *The Front Bottoms's*, *Manchester Orchestra*, or *All Get Out* contributed the most to the collaboratory song *Allentown* (Ross, 2018). This week we completed our third lab dealing with this question. For this week's lab we manipulated and used the data we collected last lab to finally come to a conclusion of which badn contirbuted most to the song.In the end we came to the conclusion that of the three bands, *Manchester Orchestra* contributed the most to the song.

Keywords: Data Analysis : Graphing : Tidyverse

1 Introduction

This lab is the culmination of the three part lab series which we have been completing for the past several weeks. Last week we acquired and organized important music data from different sources to help us determine which band contributed the most to the song in this lab. This week, through analyzing the data collected in our prior labs we were able to come to the conclusion that *Manchester Orchestra* contributed the most to the song *Allentown*. Throughout this lab, we utilized the `stringr` (Wickham, 2023), `jsonlite` (Ooms, 2014), and `tidyverse` (Wickham et al., 2019) packages to complete the majority of our tasks. We also utilized both `ggplot2` (Wickham, 2016), and the `Shiny App` provided to us via `The Data Science Collaboratory at Colgate University` (The Data Science Collaboratory at Colgate University, 2024) to create all graphs seen later on. This lab report will go through how we analyzed our data, and how we used this analysis to make our final determination that *Manchester Orchestra* contributed the most to the song.

2 Methods

For this lab, we began by loading in the `Essentia` (Alonso-Jiménez et al., 2020) data which we collected last lab. Our first task to analyze this data was to use `tidyverse` (Wickham et al., 2019) to create a function which we could use to determine whether the song `Allentown` is out of range, unusual, or in range in regards to each band's catalog of songs. The first feature of our function used the `summarize()` function from `tidyverse` to calculate the minimum, lower fence, upper fence, and maximum values that each band's catalog had for every feature in our `Essentia` Data set. We then used the `mutate()` function from `tidyverse` to create three new columns, those being *out.of.range*, *unusual* and *description*. These new columns aimed to compare the values we calculated for every feature for each band's catalog to the values of those same features for *Allentown*.

Specifically, *out.of.range* would come back as **TRUE** when the given feature's value for Allentown was less than the minimum value or more than the maximum value for that same feature in relation to each band, and would come back as **FALSE** otherwise. *Unusual* would come back **TRUE** when a given feature's value for Allentown was less than the lower fence (LF) or more than the upper fence (UF) for the given feature for each band, and would come back as **FALSE** otherwise. Finally, *description* would come back as **Out of Range** when *out.of.range* was TRUE, would come back as **Outlying** when *unusual* was TRUE, and would come back as **Within Range** otherwise.

Once we had all of this completed, we were able to run our function through a for loop which ran through every `Essentia` feature in our data set. We then filled an empty *tibble* we created with all of this data. I also decided to use `mutate()` once again to create a column which kept track of which feature each row of data was for. When running our loop, we decided to eliminate all columns from our `Essentia` data which had non numerical data. These columns ended up being *artist, album, track, chords scale, chords key, key, and mode.*

Next, we were able to go through our new tibble full of data and pick out specific features that would be useful to determining which band contributed most to the song. For this step, I specifically chose 10 features where the *description* for one band was **Within Range**, but the *description* for the other two bands were either **Outlying** or **Out of Range**. I chose to pick my specific features to analyze this way because if one band is in range to *Allentown* and two are not, it stands to reason that the band in range had more of an effect on the song.

To conclude, we created a LaTeXtable that summarized our selected features we used to determine which band contributed most to the song. This table can be found in the **Appendix** section. We also finished off by creating a couple of graphs using both `ggplot2` and the `Shiny App`, some of which will be in the **Appendix** and some will be below in the **Results** Section.

3 Results

Our final data frame ended up consisting of 181 rows and 140 columns. This means that for all 181 songs, we have 140 different types of data about each song coming from our various data sources. We can verify that no data was lost when merging since our data frame of `Essentia streaming music extractor` data had 181 rows and 11 columns, our `LIWC` data had 181 rows and 121 columns, and our `EssentiaOutput` data had 181 rows and 14 columns. When merging these three data frames together we would expect to see 181 rows and 146 columns, since the number of rows will not change cause each row corresponds to one song and each data has the same set of songs. Our data only has 140 rows since we merged by "artist", "track", and "album". This means that these columns were not duplicated since they were the same in each data frame, therefor, it makes sense that we are 6 rows short of what we expected because two sets of these 3 columns were taking out. So, we correctly have 140 columns and 181 rows showing that no data was lost during our data altering or data merging steps. Viewing our final data frame there are also no visible issues concerning the naming of columns or data within the data frame itself. All of our data appears to be present with no "NA's", and no columns appear to be misnamed. Concerning the graphs we made, I will place the one which I believe has the highest possibility of being relevant in the appendix section below.

4 Discussion

The graph I chose to include is a Violin Plot created with the *Shiny App*. For each artist, this plot shows how the happiness of their catalog of songs can be distributed. Our data on happiness came from the `EssentiaOutput` data set. Looking at the graph, it appears as though the happiness level in *Manchester Orchestra's* catalog most closely alligns with the level of happiness in the song which *The Front Bottoms* and *Manchester Orchestra* created together. So, this graph provides some evidence that possibly *Manchester Orchestra* contributed most to the song. However, in our next lab we will go into much greater detail on analyzing and visualizing our data.

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, E., Gulati, S., Herrera, P., Mayor, O., Roma, G., Salamon, J., Zapata, J., and Serra, X. (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. ISMIR.

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

Pennebaker, J. W., Boyd, R. L., Jordan, K., and Blackburn, K. (2015). *The Development and Psychometric Properties of LIWC2015*. LIWC.net.

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

The Data Science Collaboratory at Colgate University (2024). Collaboratory resources. Online Application. Retrieved from https://www.colgate.edu/about/campus-services-and-resources/data-science-collaboratory.

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

	feature	artist	out.of.range	unusual	description
1	spectral_skewness	All Get Out	FALSE	TRUE	Outlying
2	spectral_skewness	Manchester Orchestra	FALSE	FALSE	Within Range
3	spectral_skewness	The Front Bottoms	TRUE	TRUE	Out of Range
4	spectral_rolloff	All Get Out	TRUE	TRUE	Out of Range
5	spectral_rolloff	Manchester Orchestra	FALSE	FALSE	Within Range
6	spectral_rolloff	The Front Bottoms	TRUE	TRUE	Out of Range
7	spectral_kurtosis	All Get Out	FALSE	TRUE	Outlying
8	spectral_kurtosis	Manchester Orchestra	FALSE	FALSE	Within Range
9	spectral_kurtosis	The Front Bottoms	TRUE	TRUE	Out of Range
10	spectral_entropy	All Get Out	FALSE	TRUE	Outlying
11	spectral_entropy	Manchester Orchestra	FALSE	FALSE	Within Range
12	spectral_entropy	The Front Bottoms	TRUE	TRUE	Out of Range
13	spectral_energyband_middle_high	All Get Out	TRUE	TRUE	Out of Range
14	spectral_energyband_middle_high	Manchester Orchestra	FALSE	FALSE	Within Range
15	spectral_energyband_middle_high	The Front Bottoms	TRUE	TRUE	Out of Range
16	spectral_complexity	All Get Out	TRUE	TRUE	Out of Range
17	spectral_complexity	Manchester Orchestra	FALSE	FALSE	Within Range
18	spectral_complexity	The Front Bottoms	TRUE	TRUE	Out of Range
19	spectral_centroid	All Get Out	TRUE	FALSE	Out of Range
20	spectral_centroid	Manchester Orchestra	FALSE	FALSE	Within Range
21	spectral_centroid	The Front Bottoms	TRUE	FALSE	Out of Range
22	erbbands_skewness	All Get Out	TRUE	TRUE	Out of Range
23	erbbands_skewness	Manchester Orchestra	FALSE	FALSE	Within Range
24	erbbands_skewness	The Front Bottoms	TRUE	TRUE	Out of Range
25	dissonance	All Get Out	FALSE	TRUE	Outlying
26	dissonance	Manchester Orchestra	FALSE	FALSE	Within Range
27	dissonance	The Front Bottoms	TRUE	TRUE	Out of Range
28	barkbands_skewness	All Get Out	TRUE	TRUE	Out of Range
29	barkbands_skewness	Manchester Orchestra	FALSE	FALSE	Within Range
30	barkbands_skewness	The Front Bottoms	TRUE	TRUE	Out of Range

Table 1: Summary of Selected Features