Lab 05 – MATH 240 – Computational Statistics

Harrison Wolfe
Colgate University
Math Department
hwolfe@colgate.edu

Abstract

The goal of this lab was to create a set of data from 3 different bands. Using this data we wanted to show which band contributed most to the song "Allentown". We will do this by analyzing the trends from other songs from these artists then comparing that to the song itself.

Keywords: Loops; Batch files; Organzing data; Subsetting; `tidyverse`

1 Introduction

This paper is intended to describe the process in which we can extract data from music and use that data to determine if the band, All Get Out, Manchester Orchestra, or the Front Bottoms contributed most to the song Allentown. There are several important components to music like Key, Lyrics, Overall Loudness, Emotion (Happiness, Sadness, Aggressiveness) and many more. Music can be summarized in many different ways and we attempt to analyze many quantifiable categories to make the most accurate possible guess. We will make this guess based on plots and tables created using the many different data points collected.

1.1 Tasks

In this lab we are dealing with 181 tracks from 3 different artists and various albums. We had to pull the data from two spreadsheets (csv files) and and set that up as a data frame with specific data points like those listed above (loudness, tempo, key, etc) as the columns. After this we had to take an imported json file from the `jsonlite` package and turn that into a spreadsheet until we eventually combined that data with the rest (Ooms, 2014). The columns displayed data about the various musical properties while the rows represented the tracks. After we had all this data compiled into one data frame we were going to use that data to show which band had contributed most to Allentown.

2 Methods

2.1 Task 1 Methods

First, we imported various information about 181 different songs from 2 spreadsheets and a json file. Each spreadsheet had different types of data about the various tracks. The first set of data came from `Essentia Extractor Data` which gave us information about the statistics of music like the overall loudness and the spectral energy (Bogdanov et al., 2013). The data from `Essentia Extractor Data` was inputted as .wav files (normal songs) and outputted as json files which the `jsonlite` package then converted into a very long list which we turned into a spreadsheet (Ooms, 2014). The second spreadsheet came from `Essentia Models` and gave us information about the mood or vibe of the song (Alonso-Jiménez et al., 2020). This described information about how happy, sad, aggressive, etc. the songs were. The final spreadhseet came from `LIWC Output` and analyzed the lyrics of each song (Boyd et al., 2022). It provided a count of certain repeated words across each song. After putting all the spreadsheets into one larger one grouped by each track we were able to keep the datapoints (columns) that were most relevant to determining which band contributed most. This was also made possible using the `stringR` package from the `tidyverse` package to organize the data (Wickham et al., 2019).

2.2 Task 2 Methods

Using the data from task 1 we then created a list of statistics about each band in each of these categories. With those statistics we were able to say whether Allentown was within the middle 50 percent (IQR) of data for each band (Within Range), whether it was not even in the range of something ever created by the band (Out of Range), or whether it as an outlier for this piece of data, meaning that it was quite far away from the IQR (Outlying). We did this for each statistics for the bands. After we had this data, we counted the total number of qualities for each band that were "Within Range", "Out of Range", or "Outlying" and graphed them.

3 Results

From this lab we were able to create a spreadsheet with an ample amount of data about artists, The Front Bottoms, All Get Out, and Manchester Orchestra. Using this data we were able to create column plots and a table based on the amount of data that was Within Range, Out of Range, or Outlying for each given statistic compared the the measurement of that statistics for the track Allentown. These column plots and table can be seen in the appendix as Table 1, Figure 1, Figure 2, and Figure 3. Each of the column plots were made using `ggplot2` and the table was made using `xtable` (Wickham, 2016) (Dahl et al., 2019).

4 Discussion

Using the data from Table 1 and Figures 1, 2, and 3 we can clearly see that the band Manchester Orchestra has the highest amount of data points "Within Range" and the least amount of data points that are both "Out of Range" and "Outlying". From this we can say that Manchester Orchestra is the band that contributed most becuase they have the most amount of categories that are similar in the acoustic, lyrical, and tonal categories. This means that Allentown would fit Manchester Orchestra best because it is the most similar in the data points that were analyzed in this lab. It is important to note that while the data from this lab leads us to believe that Manchester Orchestra contributed the most there is always a possibility that the other bands contributed more. Speculation based on the data is the best we can do in this scenario however since Manchester Orchestra has so few categories that are "Out of Range" and "Outlying" it is logical to assume that this could be very well be their song that they contributed most to.

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.

Bogdanov, D., Wack, N., Gómez, E., Gulati, S., Herrera, P., Mayor, O., Roma, G., Salamon, J., Zapata, J. R., Serra, X., et al. (2013). Essentia: An audio analysis library for music information retrieval. In *ISMIR*, volume 13, 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*, 10:1–47.

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.

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

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

These graphs could not fit in the template above so they have been placed in the Appendix instead. They are tables and column graphs demonstrating the data from above. These tables and graphs were made using `xtable` and `ggplot2` respectively (Dahl et al., 2019) (Wickham, 2016).

Artist	Within Range	Out of Range	Outlying
All Get Out	158.00	22.00	17.00
Manchester Orchestra	183.00	3.00	11.00
The Front Bottoms	156.00	30.00	11.00

Table 1: Table that shows how many datapoints are Within Range, Out of Range, or Outyling for each band

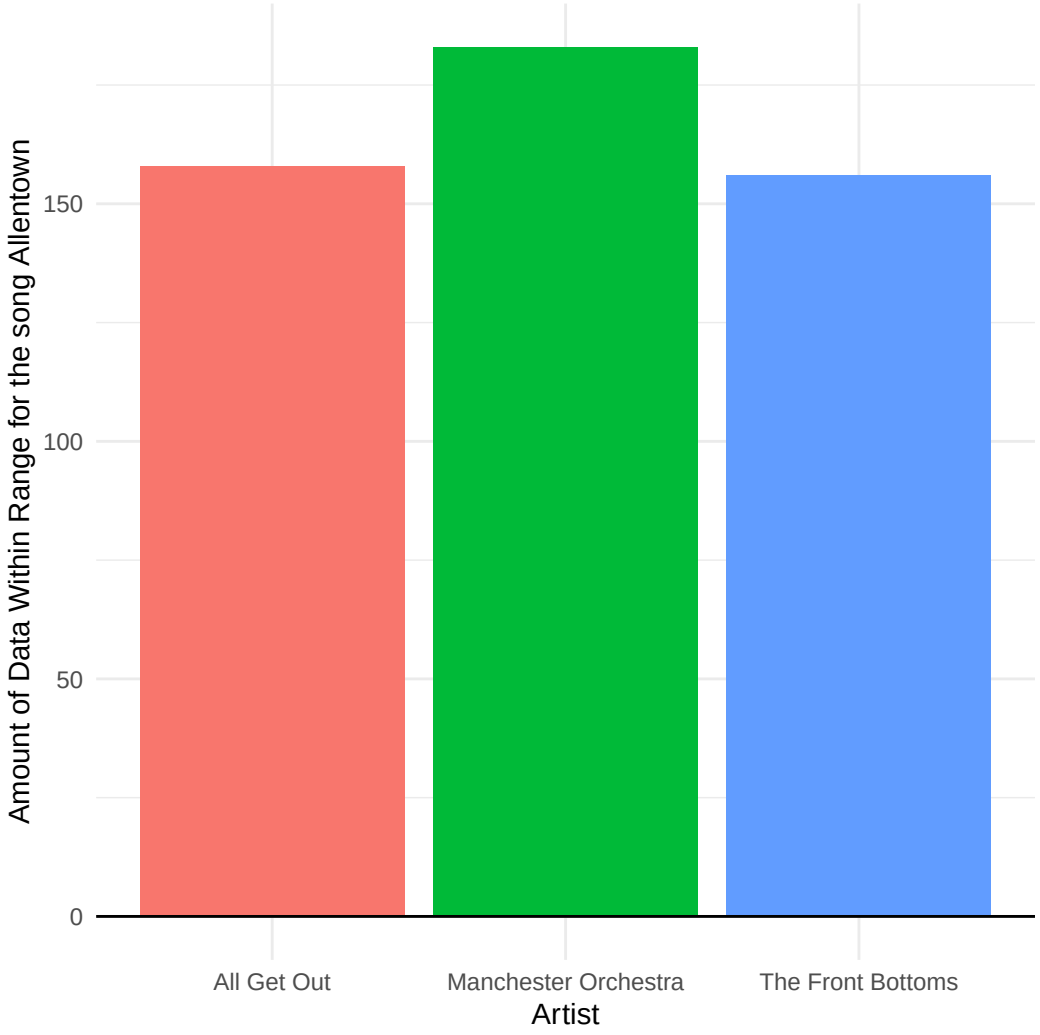

Figure 1: Number of Statistics Within Range for Each Band

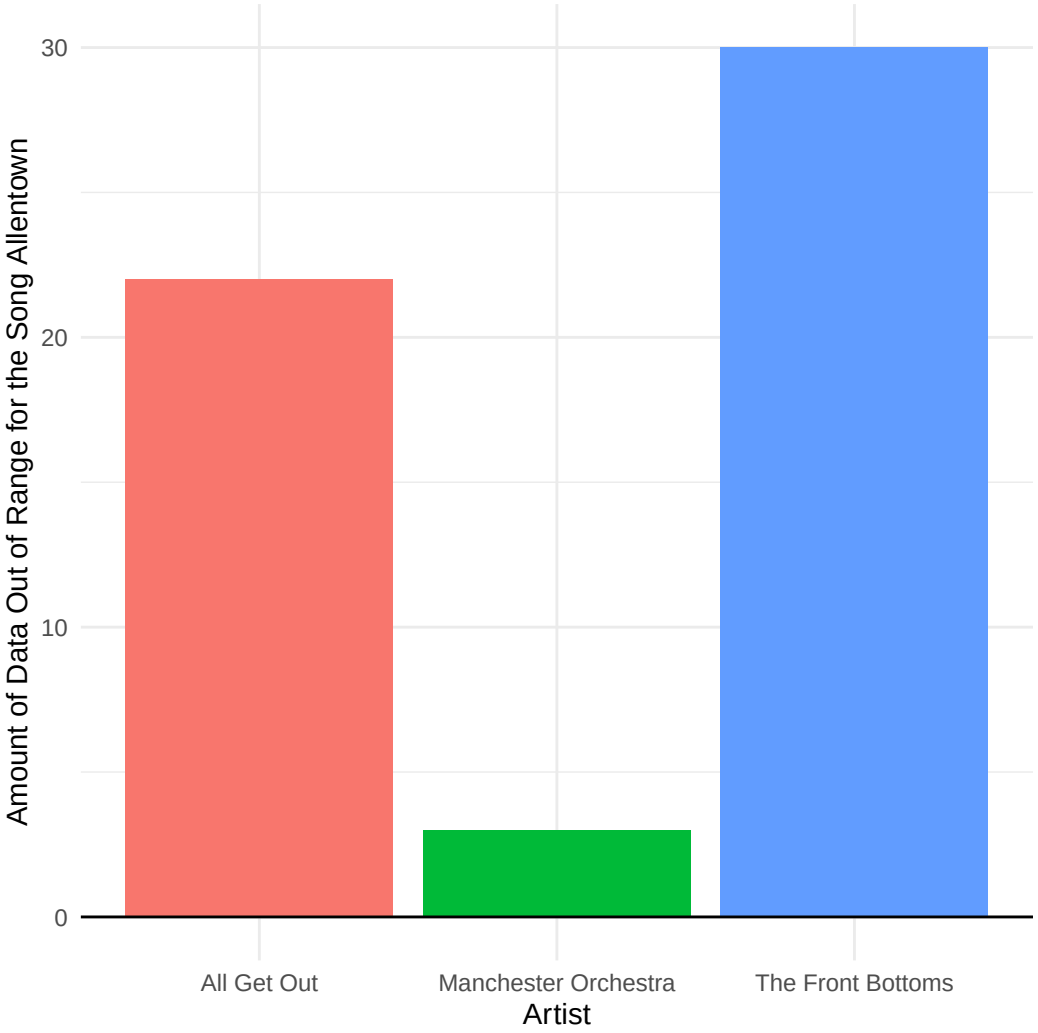

Figure 2: Number of Statistics Out of Range for Each Band

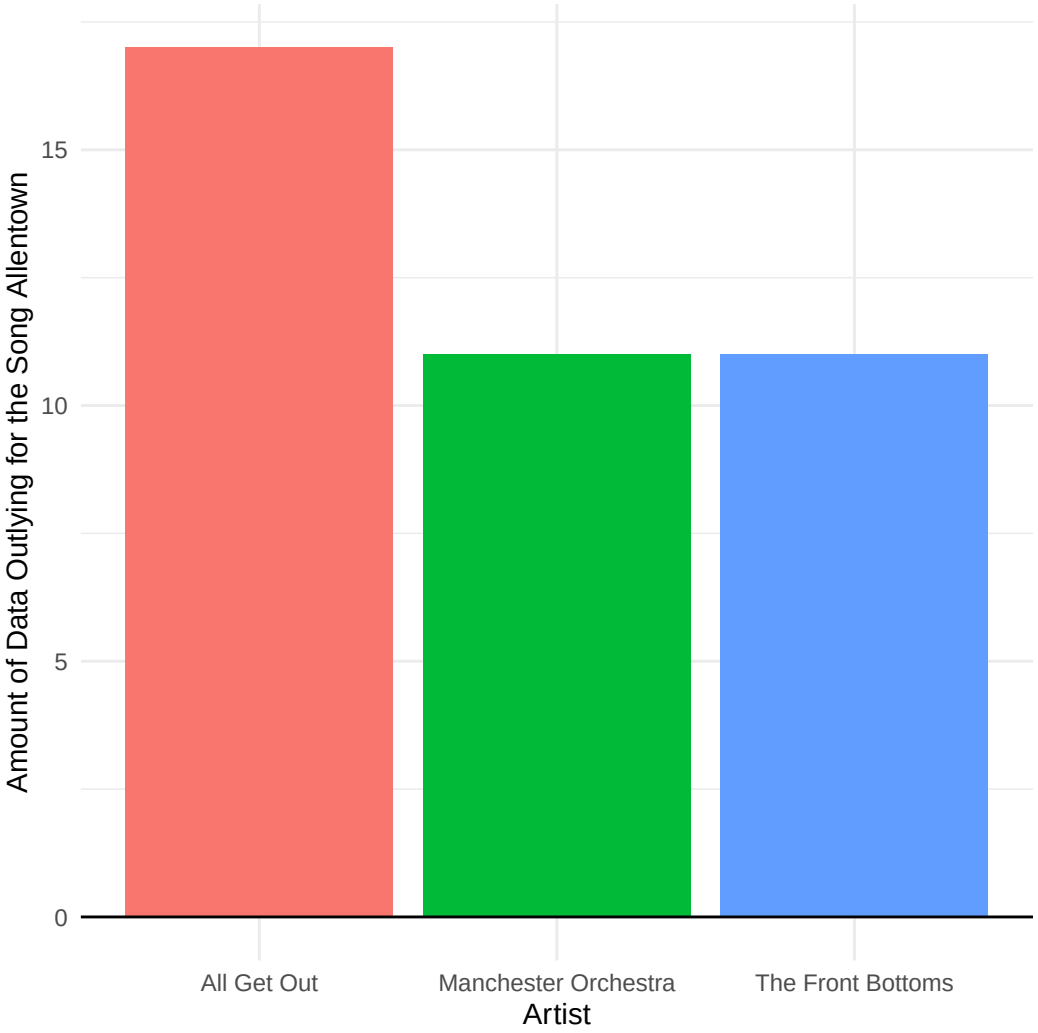

Figure 3: Number of Statistics Outlying for Each Band