Final Project Report Team No. 6

Ritika Agrawal, Bernard Fay, Tyler Smith, and Chase Weida

Our dataset covered the top 100 albums in nine different genres as rated by Rate Your Music fans online. After obtaining this data, we got the Spotify API codes for each album, which then translated into variables for each album that we could analyze. There were four main questions covered in our analysis. These included how labels varied between decades, how labels influenced the correlation between danceability and valence, how labels influenced genres and artists, and comparing metrics between genres.

We decided very quickly that we would like to use a pairing of Spotify API and a music-oriented dataset to shed light on the structure within the dataset. Our search for data began on traditional data repositories such as Kaggle and UCI. Unfortunately, there were no datasets which piqued our interest. Instead, we came upon a series of RateYourMusic.com lists which encompassed a total of 900 data points across nine genres. These lists were stored on nine separate web pages in plain text. To acquire the data, we wrote a function in R which accepts the webpage url and parses the artist name, album name, and ranking within its genre. In the process of obtaining the information from RateYourMusic.com, we encountered slight roadblocks from the site's firewall which blocked us from accessing the website at a rapid pace through R.

Fortunately, once the information was pulled and parsed it could be compiled into a single dataframe and written to a csv file that would allow us to access the information while circumventing the need to excessively access the website. From here, we needed to pair each album and artist combination with their respective musical metrics from Spotify. Since the Spotifyr package does not support vectorized calls to the Spotify API, we decided that it was better to take our dataframe and load it into Python to utilize the Spotipy package. We queried the API with the artist name and album title to find the album ID for each of the albums. With the Album ID, accessing the musical metrics for each song on every album was completed with a quick search. Now that the data frame had been augmented with our eight musical metrics plus a few other points of album metadata, we wrote it again to a csv and imported it into R to continue with data cleaning and analysis.

Once we had the data, we constructed a method to ensure that everybody's individual skill was given a chance to flourish. Independent from each other, we all explored the data through our own lens. With a host of takeaways from initial exploration, we had analysis on a wide scope of variables and genres. This analysis shaped the base for our four questions of interest, and eventually led to the final decision to analyze the overall effect of labels in a majority of the analysis.

Before diving directly into the musical metrics, we decided to assume a wider point of view and analyze the relationship between genres and labels from the perspective of time. Genres were the main delineating factor of the original dataset, so we first looked at a visualization of the release year of each album by genre. This took the form of a series of boxplots depicting the quartiles of the years in which each genre's albums were released. This led to some surprising and some mundane realizations. For example, it became quite obvious that jazz, progressive rock, and soul had, on average, the oldest release dates of all albums. This was not surprising to us as these genres have not been in the mainstream spotlight for many decades. However, more surprising was the distribution of the popular rap albums. The middle 50% of rap records are set squarely in the 90s. On the one hand, the 90s saw the mainstream revolution and popularization of rap music, but one could also argue that rap music underwent important positive changes in the early 2000s that would warrant a greater proportion of representation in the list.

Next, we looked at the most prolific musical labels and their representation by genre and decade. The most prolific of the labels, Columbia Records and Columbia/Legacy, was constantly releasing popular records (about 5+ per decade). Much more surprisingly, we saw several record labels that had performed exceedingly well in singular decades but lacked presence in others. For example, EMI and Pink Floyd records in the 1960s and 1970s. Upon closer investigation of this phenomenon, we found that the Beatles and Pink Floyd were voted as very popular multiple times across the lists. Thus, EMI and Pink Floyd records success can be attributed not to the large number of popular acts on the label, but the sweeping success of only a couple of bands.

Moving on to our featured graphic analysis, we analyzed the relationship between valence and danceability. Valence describes how positive and happy-sounding the music is, while danceability describes how easy it is to dance to the music based on tempo, rhythmic stability, beat strength, and overall regularity. The strong, positive correlation between the two variables as shown in Figure 1 seems rather intuitive; it's easier to dance to uplifting music and uplifting music tends to have a regularity to it. For the sake of readability, we focused on albums from labels that produced seven or more albums in a single genre.

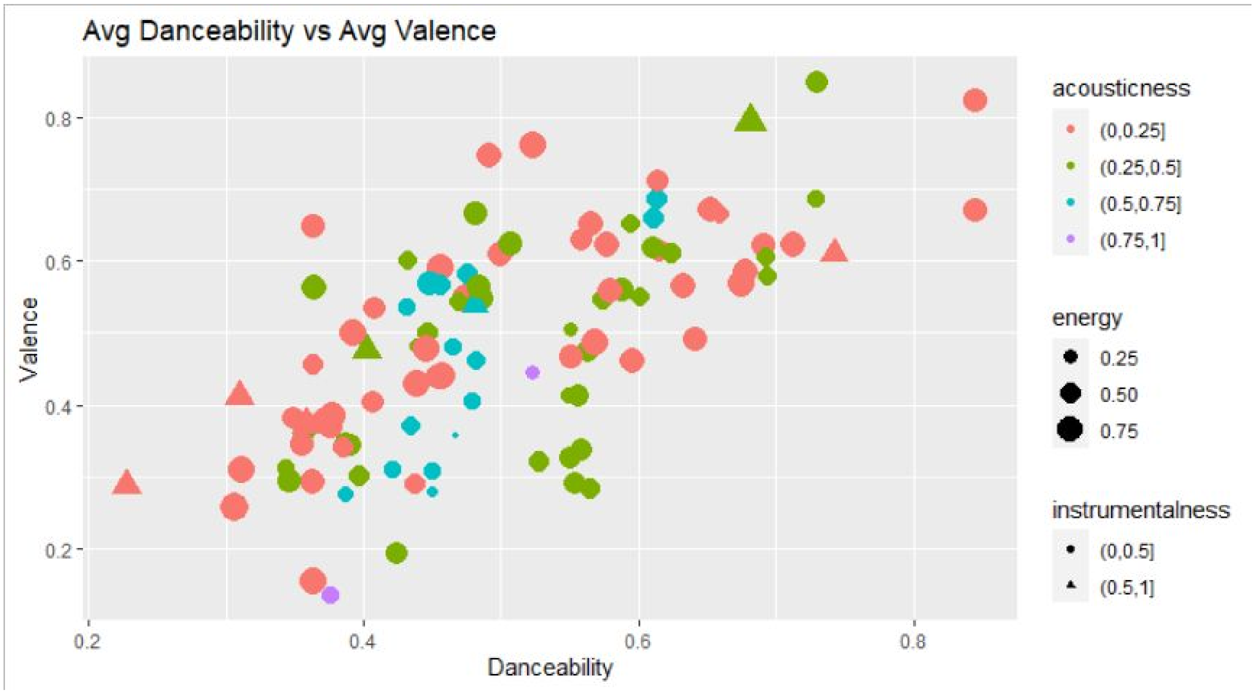

Figure 1: Featured Infographic on average Danceability vs. average Valence

Additionally, we found that there were three other variables associated with this relationship: acousticness, energy, and instrumentalness. Acousticness attempts to quantify how much electronic amplification is present with 1 being amplification throughout and 0 being no amplification. Energy describes the intensity of the music, with higher values corresponding to higher energy. Instrumentalness identifies how vocals are used in the album. Higher values indicate little to no vocals, with lower values equating to more vocals. Albums with an instrumentalness of 0.5 and higher are defined as instrumental tracks by Spotify.

Figure 2 displays that low acoustic songs tend to gravitate toward the extremes of the danceability-valence relationship, while higher acoustic songs were concentrated more toward the middle of the relationship. A similar relationship can be seen with energy and the danceability-valence correlation. The overwhelming ratio of circles to triangles shows that there are relatively few albums classified as instrumental. We then decided to look at how labels factored into all of this. As mentioned earlier, we limited the data to albums from labels that produced seven or more albums on a single genre. The result was the 10 labels seen in Figure 2.

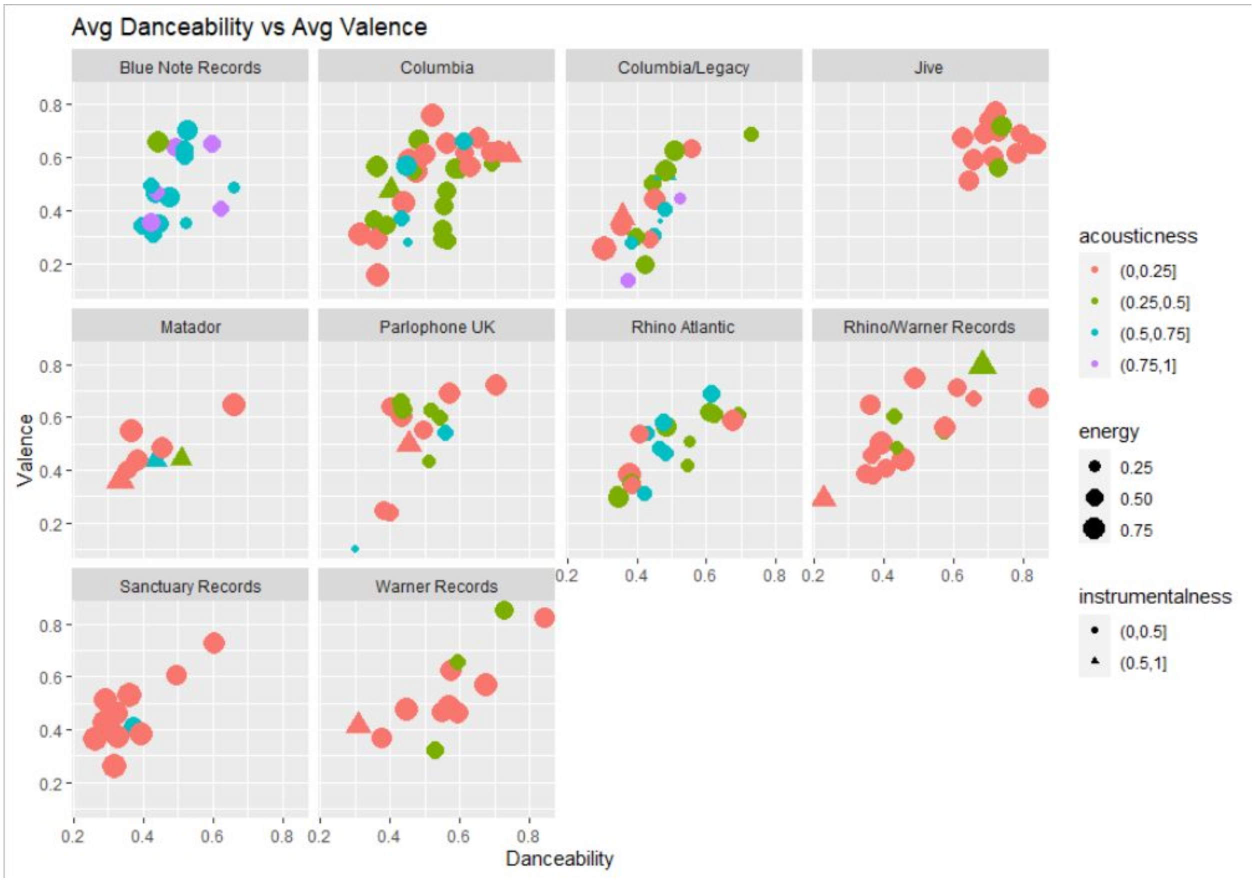

Figure 2: Danceability vs. Valence in top record labels

One of the things that immediately stands out in this plot is the concentration of points for some labels. Jive, for example, has only high-danceability, high-valence albums. Additionally, these albums tend to use a lot of electronic amplification, have high energy, and be non-instrumental tracks. Jive's focus on the pop genre aligns with this concentration as each of these characteristics is often associated with pop music. Other labels, such as Matador and Sanctuary Records, tended to focus on low-danceability, low-valence albums that had high energy and low acousticness, With a few exceptions. Matador had a close-to-even split of instrumental and non-instrumental tracks, while Sanctuary Records produced only non-instrumental tracks. In Matador's case, primarily producing metal albums can be attributed to the data concentration as well as the instrumental/non-instrumental split. Lastly, Blue Note Records produced music that was concentrated toward the center or middle of the danceability-valence relationship. Their albums tended to be acoustic in nature, middling to low energy, and non-instrumental. All of these characteristics are associated with popular jazz albums, which is exactly what Blue Note Records focused on.

Other recording labels such as Columbia, Columbia/Legacy, Rhino Atlantic, and Rhino/Warner Records produced more diverse albums. Specific concentrations aren't easily identifiable for these labels, which is likely due to these labels being some of the more prominent labels in the music industry. As such, these labels often adjusted their concentrations based on what was popular at the time, which leads to a very diverse set of albums as a whole. If each of these labels was broken down by decade, we would likely see concentration corresponding to the popular music genre at the time. We can see that labels tend to concentrate in specific parts of the danceability-valence relationship along with accompanying characteristics of acousticness, energy, and instrumentalness. Bringing these labels together into one graph gives us the relationship we saw in the first graph.

Labels appeared to also influence different artists and genres. Within the pop genre in particular, we looked at the top three pop labels in terms of the overall number of albums inside the top-100. These three labels are Columbia, Jive and Warner Records, with their albums by valence and loudness all displayed in Figure 3.

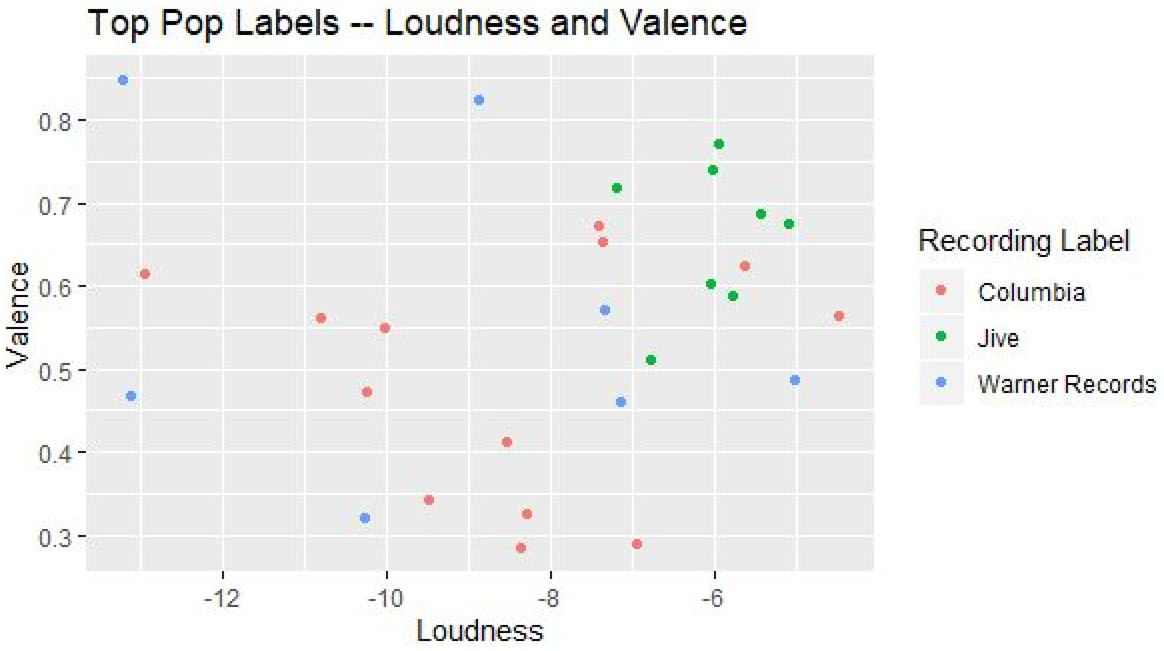

Figure 3: Loudness and Valence in top three Pop labels

Valence is a measure of happiness in the albums described earlier, while loudness is the measure of how loud the album was recorded at. This visual revealed that Jive recorded louder and in general, at a higher valence than Columbia. In detail, six of Columbia’s albums had a lower valence than Jive’s lowest-valence album. Contrasting the other two, Warner Records had albums all over the place, some resembling Columbia and others more like Jive. However, they did venture into a different territory by having the two highest-valence albums, but both were recorded at a low loudness. Figure 3 revealed that Jive and Columbia each have a distinct method with their loudness and valence in top albums.

Diving into analysis on how a label impacts a given artist, we found that jazz artist John Coltrane recorded albums with four different labels. We looked at how his work differed in valence and energy between the four labels, summarized in Figure 4 below.

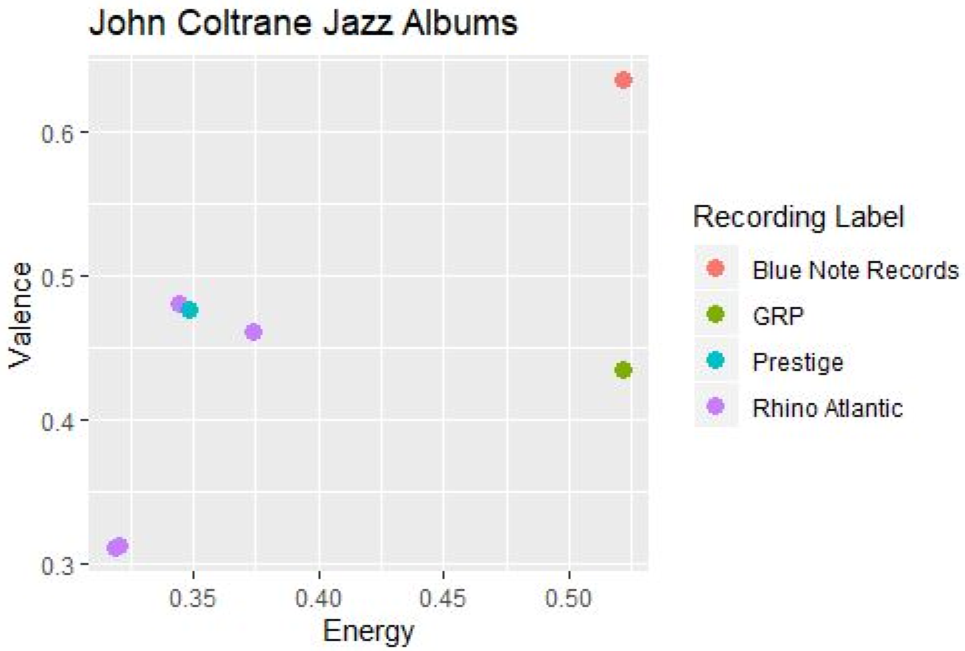

Figure 4: John Coltrane's Jazz albums by label

The plot revealed that Blue Note Records prioritized having high valence and energy, GRP had high energy and average valence, and both Rhino Atlantic and Prestige were on the lower end of energy and valence. It was very interesting to see that each label requested something different than the others from Coltrane. Each artist has their own style, but even with the same artist, the albums can widely vary. This analysis revealed the disparities between labels with the same artist.

In our final piece of analysis, we looked at metrics across genres. For example, we see from Figure 5 below how valence varies across genres.

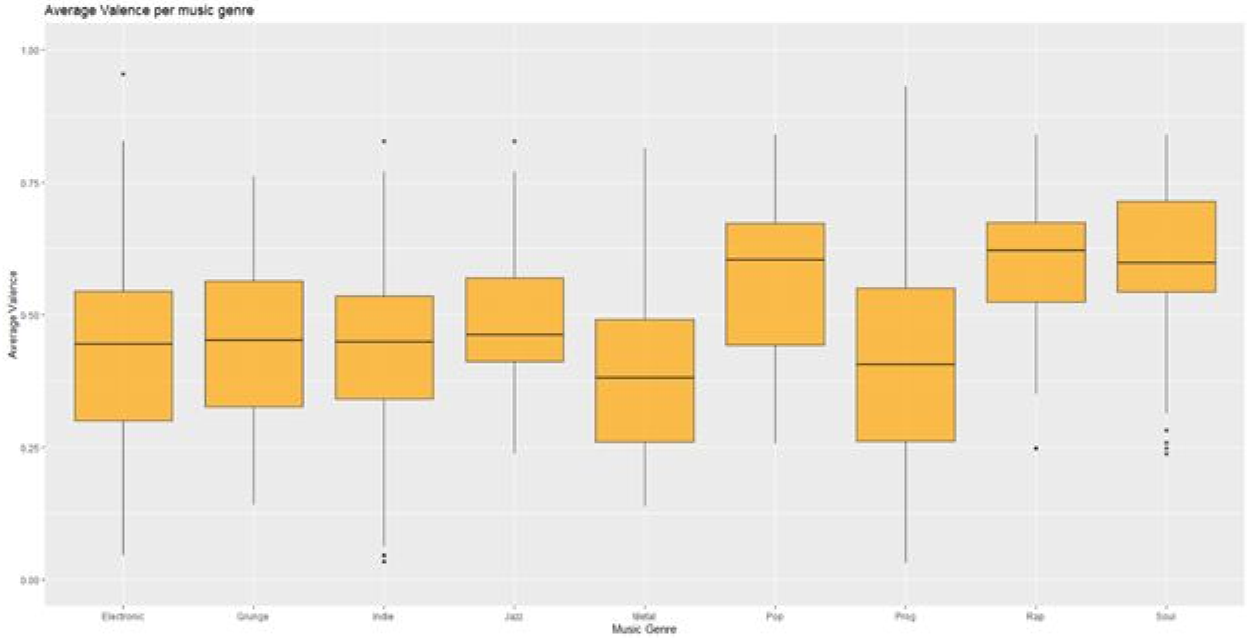

Figure 5: Average valence by genre of album

The average valence of the Soul genre is the highest with few outliers on the lower side. Average valence of Pop & Rap genres are almost the same with Pop genre having more variability. The other genres seem to have more or less similar average valence. Another worthwhile note is that progressive rock has the widest interquartile range (IQR), signifying that it is not well defined by the valence metric. Rap had the smallest IQR of all the genres, with just one outlier on the low side of valence that may have made this range wider.

We were also curious as to how valence varied from top artists across genres. Figure 6 reveals that Madonna's songs have the highest average valence followed by the Prince and Talking Heads. These three artists not only had the highest marks, they also more than doubled the next closest artists as far as average valence.

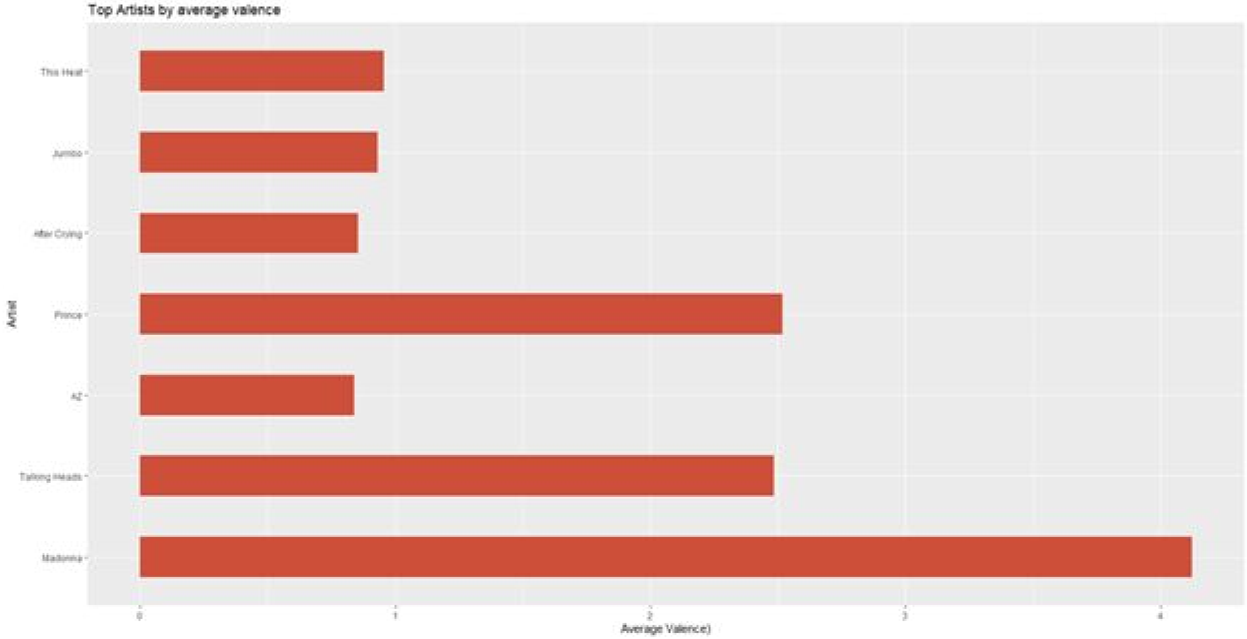

Figure 6: Top artists by average valence

The other metric of interest for this analysis was the tempo or speed of the song, displayed at the genre level in Figure 7. Average tempo of the metal genre is the highest with no outliers present. The rap and jazz genres have lower average tempos. The other genres are between 100 and 130. Overall, there did appear to be multiple outliers present. This is fascinating since this implies that tempo really does not define a genre. An artist will record at whatever tempo they would like to in order to produce their album.

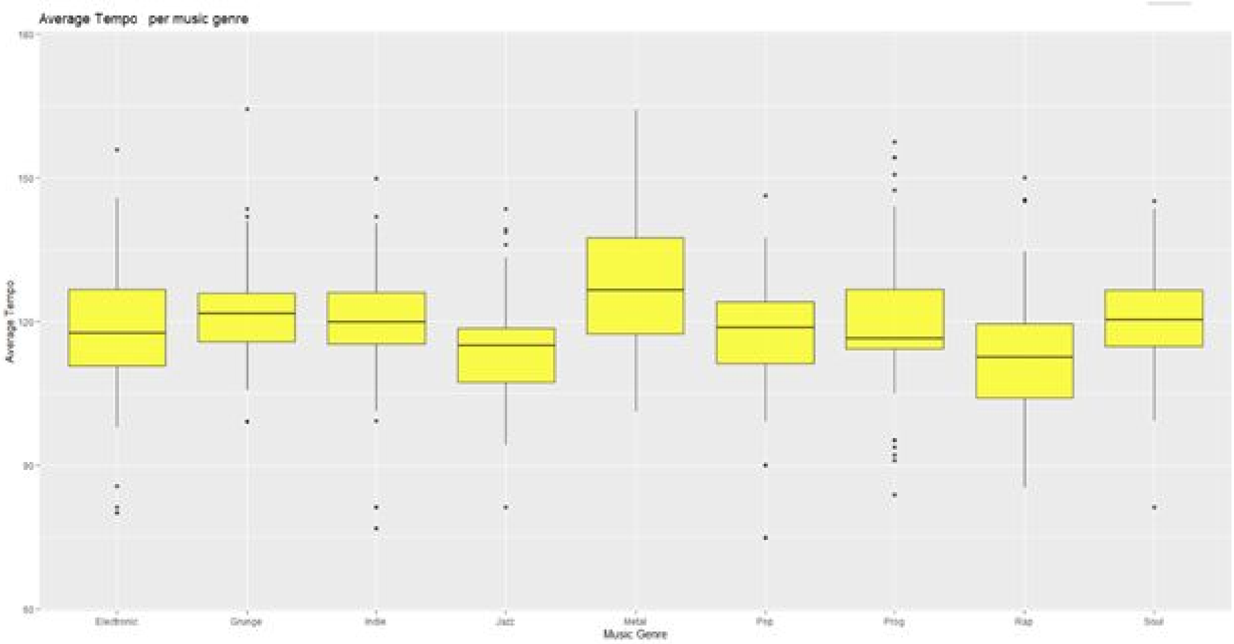

Figure 7: Average tempo by genre

Individual analysis on tempo again revealed some of the artists recorded at incredibly fast tempos comparatively to other top artists. Figure 8 below displays the fastest recording artists and Camel, Santana and Deep Purple all had extremely fast tempos in their top albums, with speeds above 300 beats per minute. No other artist had a speed above 200 in the figure.

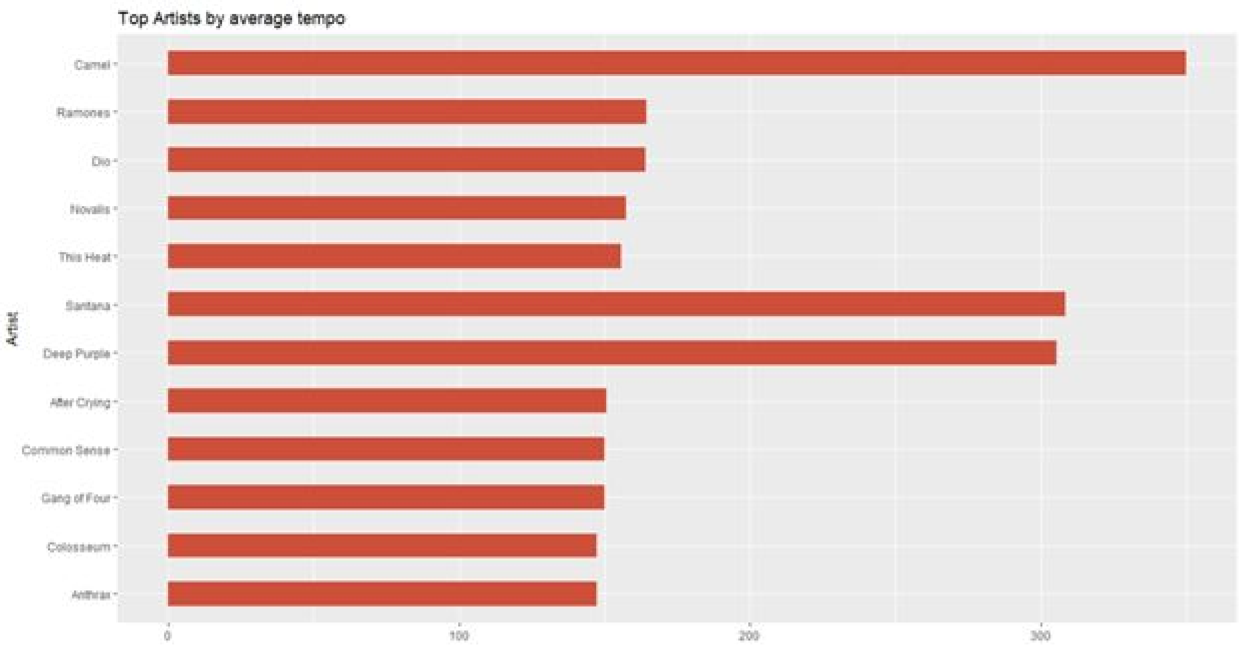

Figure 8: Top artists by average tempo

Overall, our team found a wide array of findings that this dataset offered. The takeaways from the analysis of the release date was something that we expected when we first selected this dataset. From there, we were able to uncover various relationships between variables, how the labels influenced genres and artists, and how various artists stood out individually among the metrics we had available. The main finding that most of our analysis pointed us to is that recording labels do have a significant impact on top artists and albums in music genres in a variety of ways. These differences may even be slight, but they are present and likely lead to the success each of these albums have found.

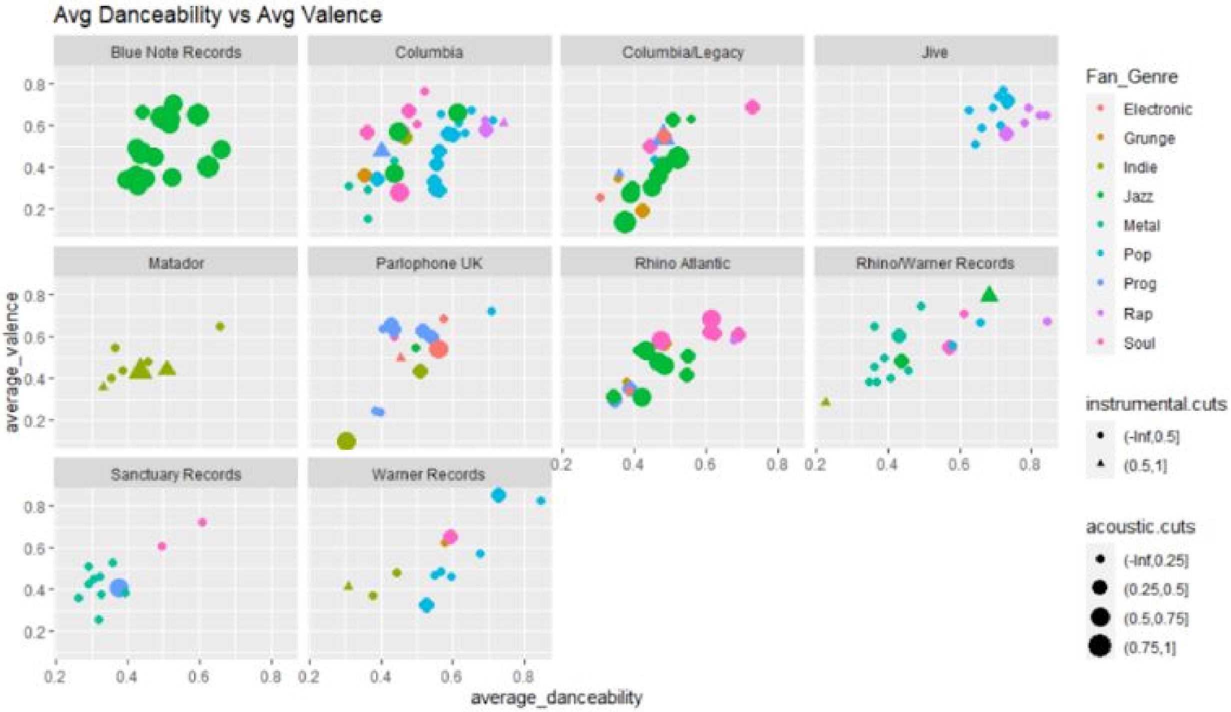

Figure 9: Labels are an important part of the music industry (Featured Infographic)