EDA Writeup

Lucy Wu

5/5/2019

We begin by exploring basic features of the dataset.

Data Descriptions

Our dataset includes all songs that appeared in the Billboard Top 100 from 2000 through 2018.

In total, we have 3,320 observations of 42 variables in our dataset. Each observation represents one song, and the variables are as follows:

Song Metadata: Non-musical characteristics of the song. - `song`: The title of the song. -`artist`: The artist that created the song. - `release_date`: The date the song was released. - `release_season`: The season (fall/winter/summer/spring) the song was released. - `release_year`: The year the song was released. - `artist.pop`: The popularity of the artist, as measured by the number of Billboard Top 100 hits they had in the three years before the song was released.

Musical Characteristics: Simple musical characteristics of the song. These characteristics are `tempo`, `mode`, `key`, `time_signature`, and `duration_ms`; each is self-explanatory.

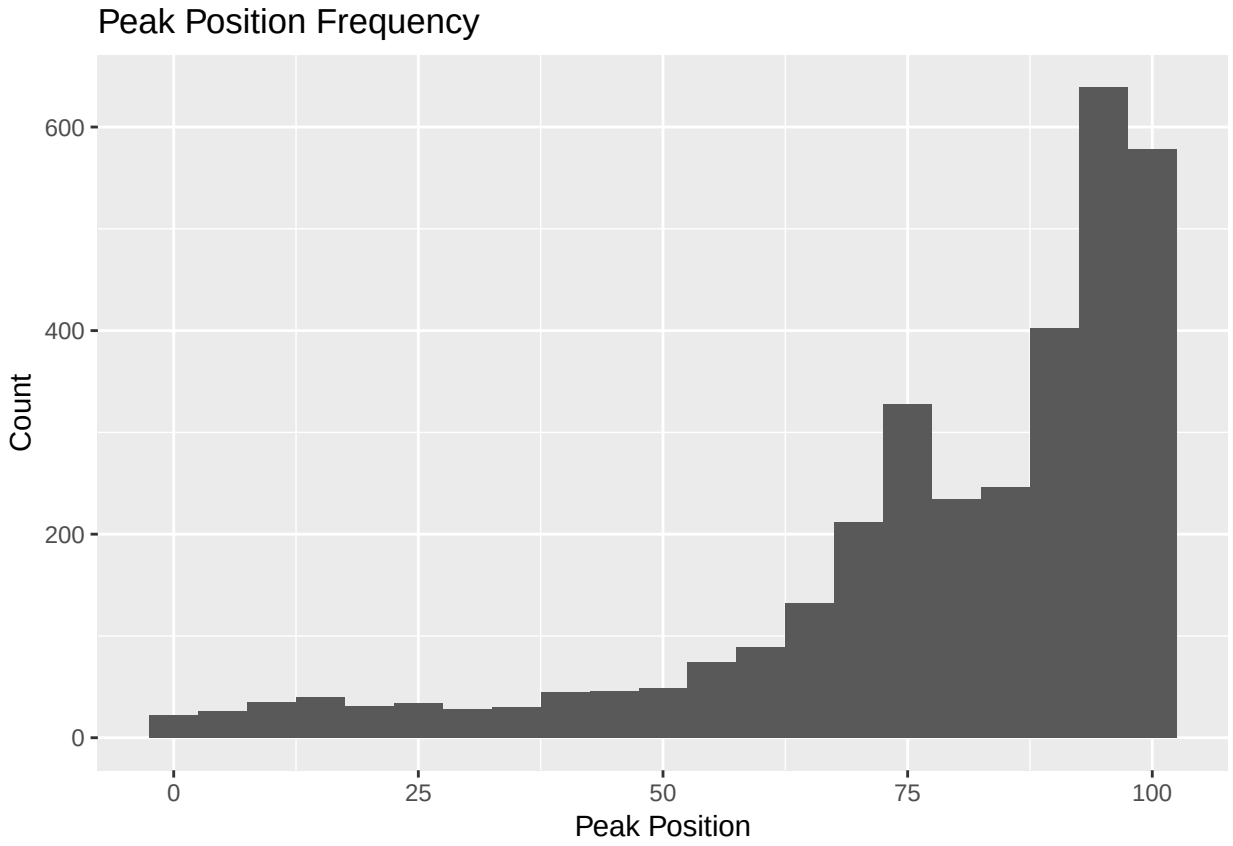
Spotify Audio Analysis Data: Complex musical characteristics of the song, as computed by Spotify. *add bibliography? - `acousticness`: How acoustic the track is (as opposed to electrically amplified). - `danceability`: How suitable the track is for dancing based on measures including its tempo, rhythm stability, beat strength. - `energy`: How intense/active the track is. For example, a Bach prelude has low energy, while Green Day's "American Idiot" has high energy. - `instrumentalness`: How likely the track is to be instrumental (have no words). For example, a symphony would generally have high instrumentalness since it has no words, while a rap song would generally have low instrumentalness. - `liveness`: How likely the track is to be a recording of a live performance. - `loudness`: Average loudness of the track in decibels. - `speechiness`: How speech-like the track is. For example, a podcast would have high speechiness, while an Adele song would have low speechiness. - `valence`: How cheerful the track sounds. For example, the Weeknd's "Often" has low valence, while One Direction's "Live While We're Young" has high valence.

Genre: Genre(s) associated with the song's artist **add footnote** according to Spotify. Since the genres provided by Spotify were extremely granular (with only a few artists in the dataset associated with each genre), we picked a few broader genre categories and categorized each song accordingly. Our broader genre categories are trap, hip-hop, indie, punk, rap, jazz, pop, metal, country, folk, bluegrass, house, rock, opera, classical, instrumental, and funk. For example, a song associated with "Philly rap" would be categorized as "rap".

Popularity (target variables): The ultimate popularity of the song. - `peak.position`: Peak position that the song reached on the Billboard Top 100 (with 1 being most popular, 100 being less popular). - `weeks.on.chart`: Total weeks that the song spent on the Billboard Top 100 chart.

Data Exploration

First, let us examine our target variable, `Peak.Position`.



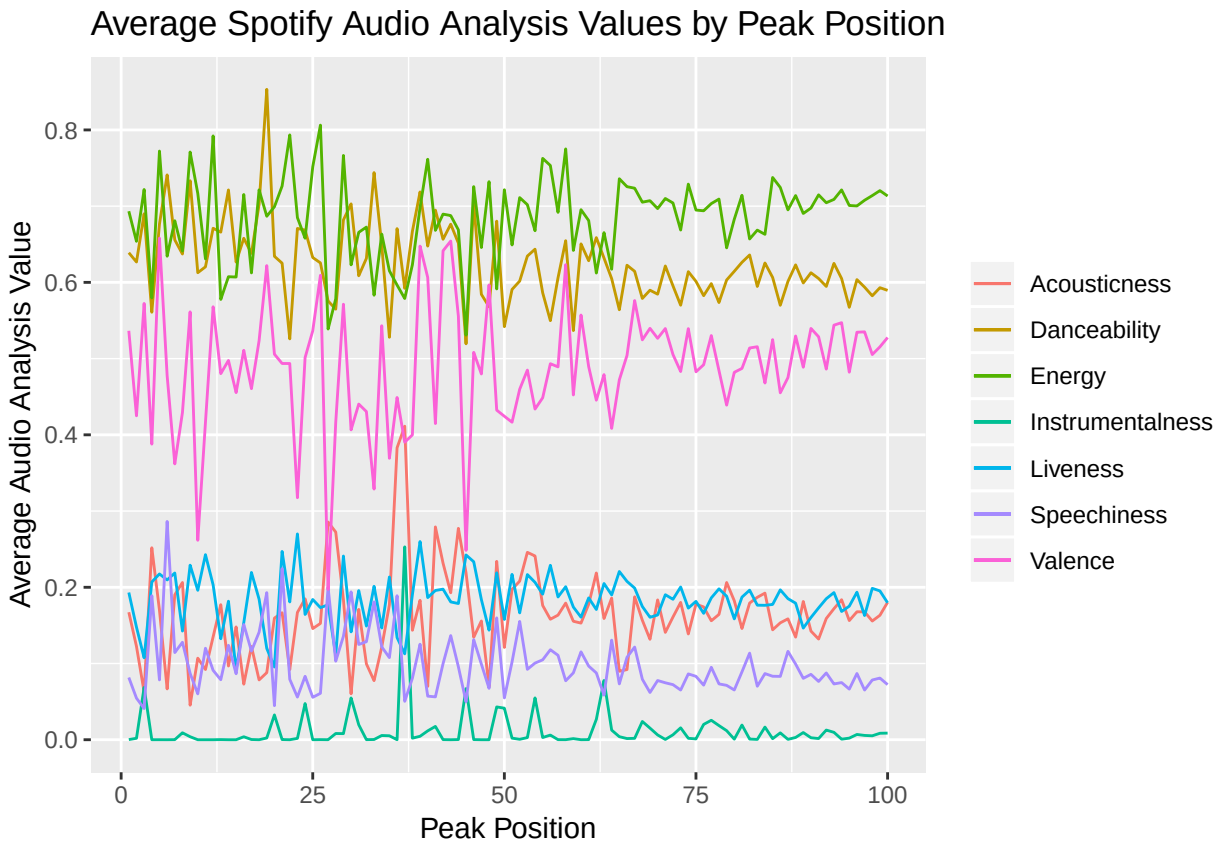
It looks like relatively few songs in the dataset reach the top 50 compared to the total number of songs that reach top 100. In other words, our dataset is unbalanced with regards to peak position.

Overall, the most popular artists are:

Artist	**Number of Top 100 Songs**
Drake	67
Taylor Swift	48
Tim McGraw	33
Keith Urban	29
Kenny Chesney	27

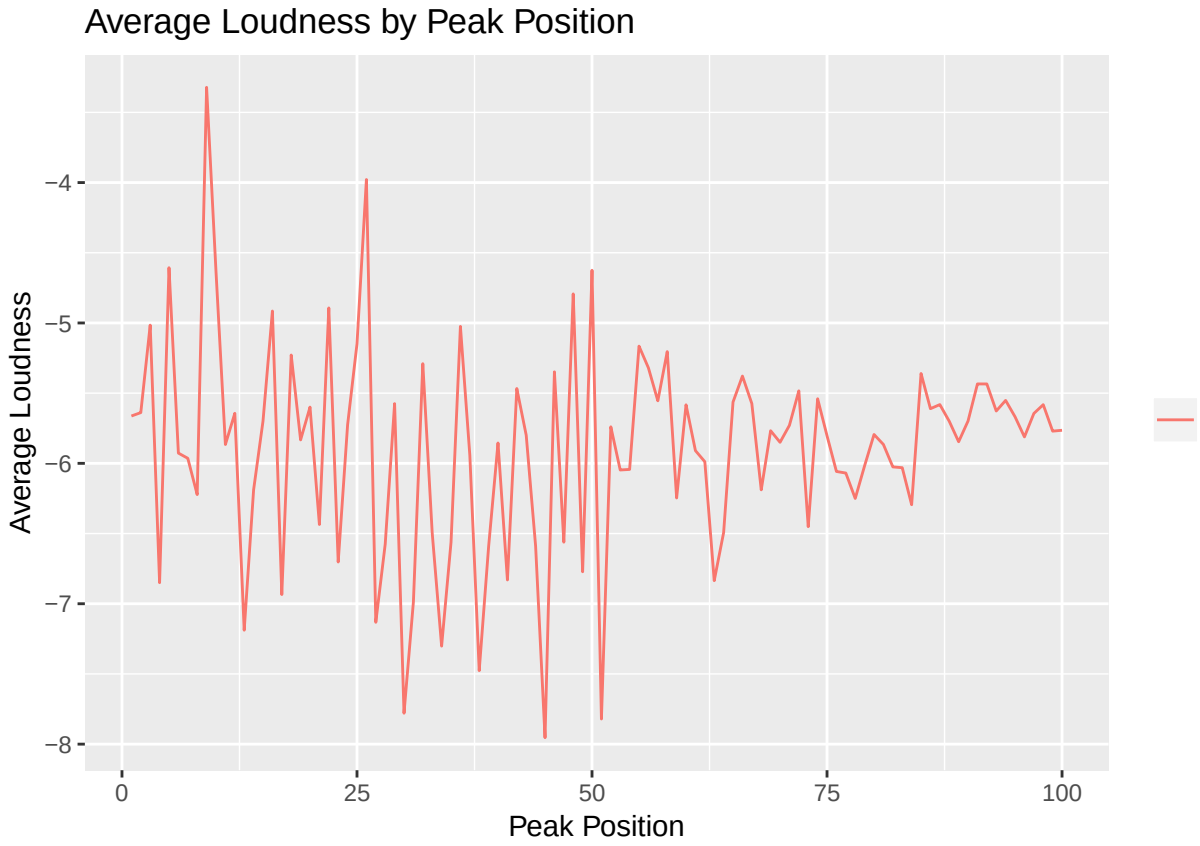
Given that we're predicting a song's popularity in terms of its peak position on the Billboard chart, we might wonder how various factors are correlated with a song's peak position.

All Spotify audio analysis variables are scaled from 0 to 1 with the exception of loudness, so we plot them together below:



As shown in the plot above, there do not appear to be strong trends relating any Spotify audio analysis to the peak position of a song; each audio analysis variable seems to take roughly constant average values for all peak positions.

We can also plot the one remaining audio analysis variable, loudness:

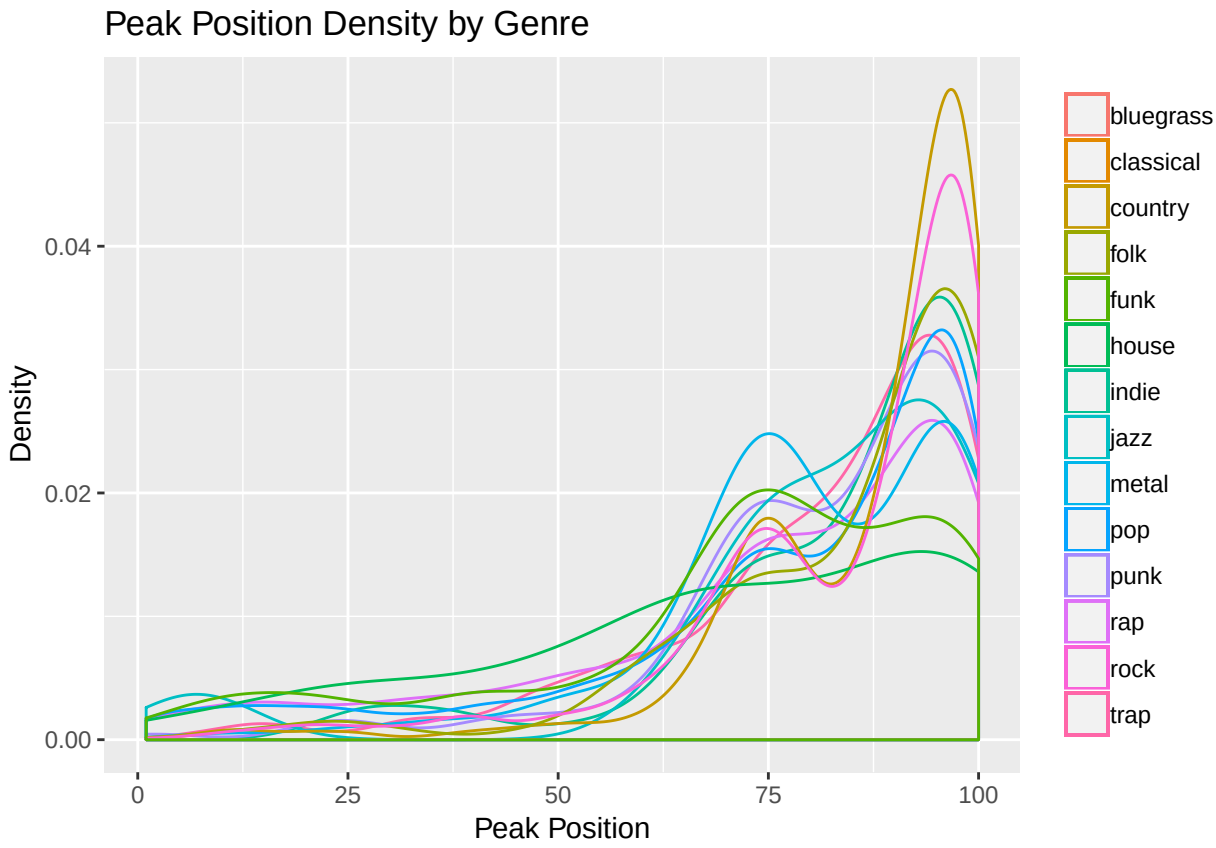


Again, there does not appear to be a clear trend relating peak position and song loudness. There is clearly additional variance in loudness among higher peak positions, but this is likely due to the fewer number of songs with high peak positions.

We might also wonder how genre plays into song popularity.

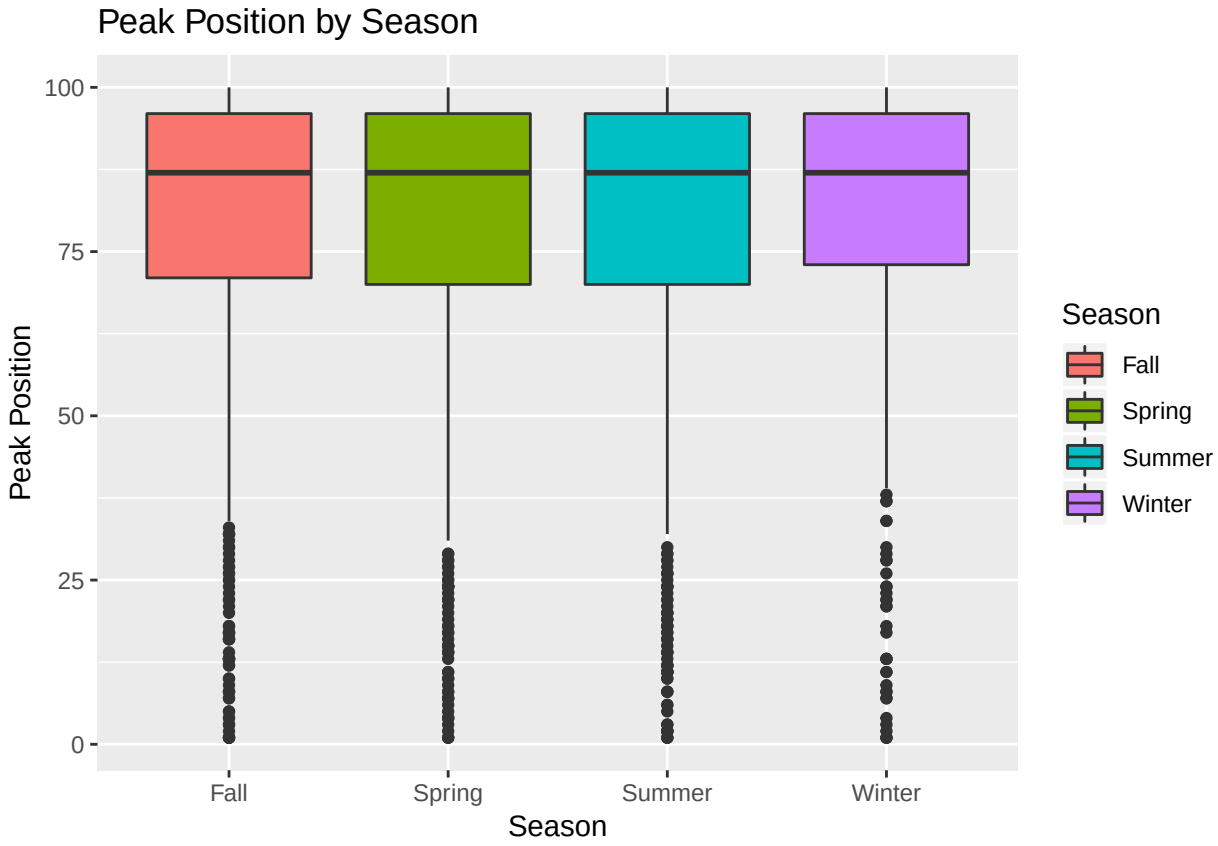
```
## Warning: Groups with fewer than two data points have been dropped.
```

```
## Warning: Groups with fewer than two data points have been dropped.
```



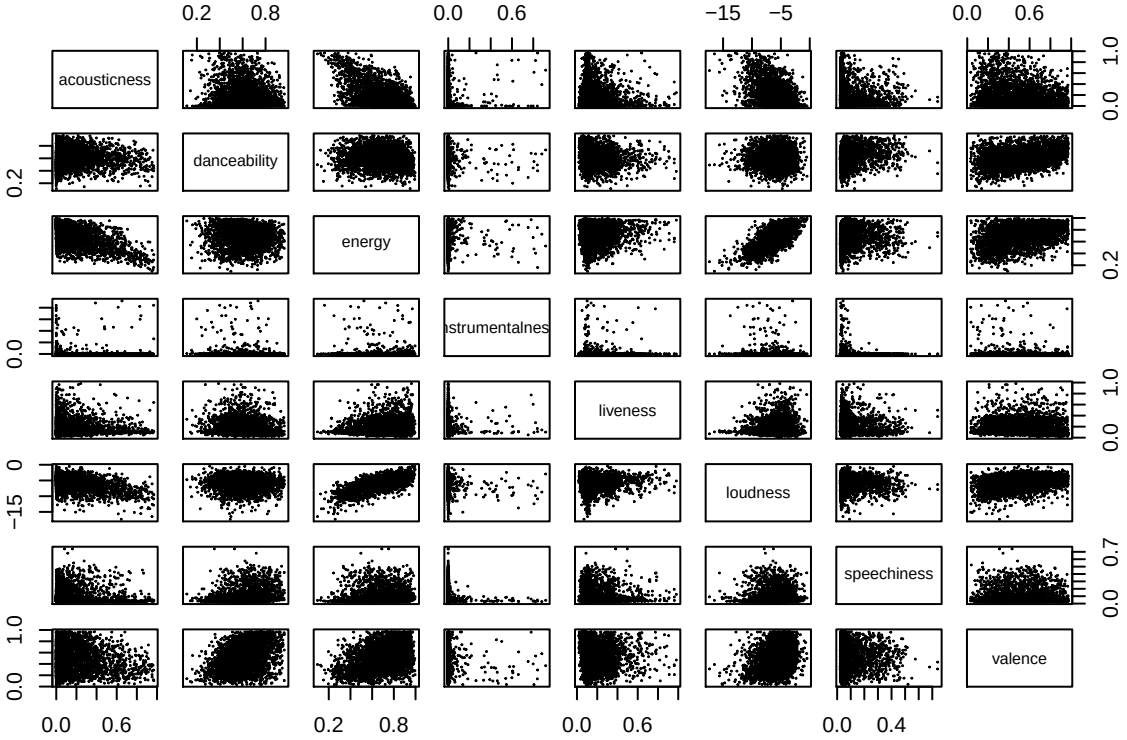
It looks like most genres are alike in that most songs within the genre peak at lower positions. However, it seems like songs associated with the rock and country genres tend to take lower peak positions compared to jazz and house music.

We can also examine how peak position changes with season.



It looks like the distribution of peak positions is roughly the same among all four seasons.

Finally, we will investigate the correlations between the variables in our dataset.

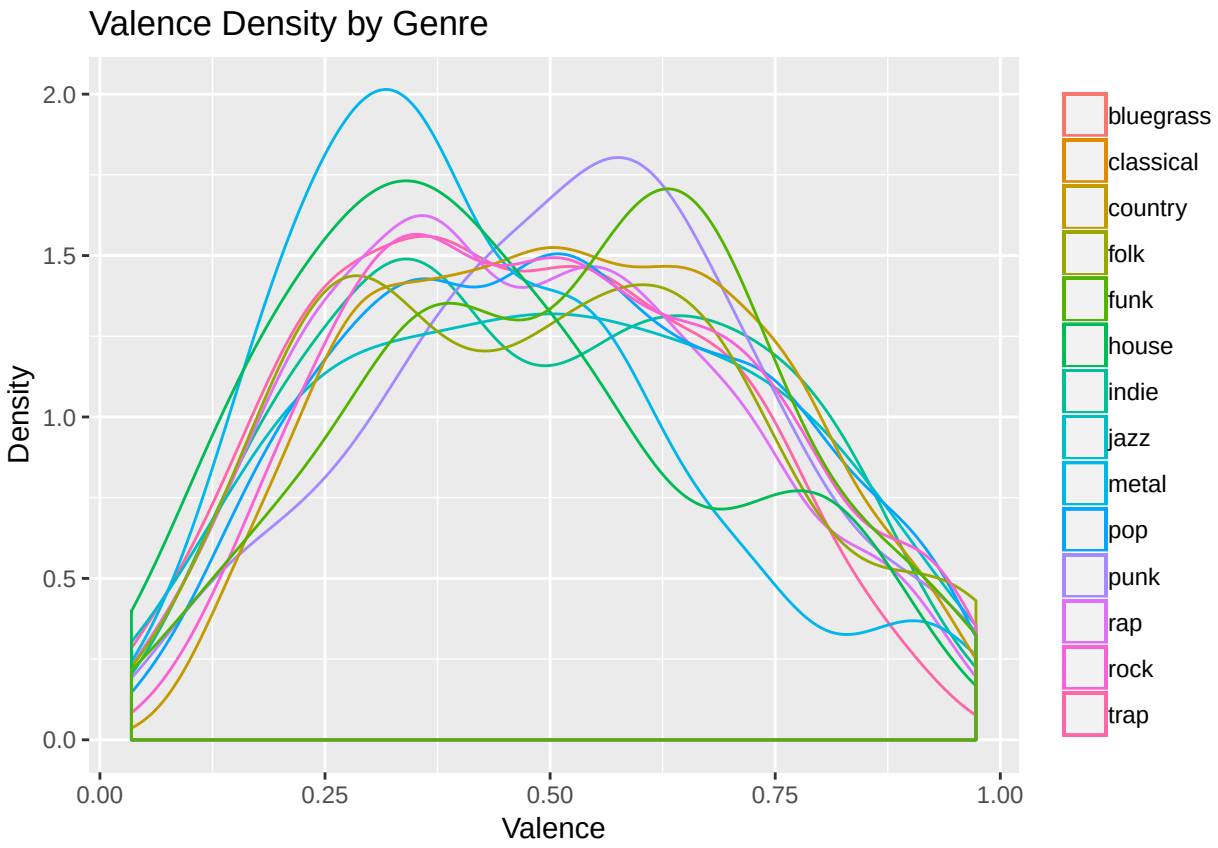


It looks like most pairs of Spotify audio analysis variables are not too strongly correlated, with the exception being energy and loudness. In particular, the correlation between energy and loudness is 0.72. Given that the correlation is relatively high, we will remove energy from the model and keep only loudness in order to preserve model interpretability.

We might also expect the Spotify audio analysis variables to be somewhat correlated with the genre.

```
## Warning: Groups with fewer than two data points have been dropped.
```

```
## Warning: Groups with fewer than two data points have been dropped.
```



Above, we see that metal tends to have low valence compared to punk, funk, and country. To conserve space, the plots relating other Spotify audio analysis variables to genre can be found in the appendix. In summary, most audio analysis variables do not seem to take notably different values among different genres. The exceptions are energy, where the punk and pop genres have much higher energy compared to indie music; and danceability, where metal takes far lower values compared to trap.

This might seem surprising, but it is likely due to the fact that the "genre" of a song in our dataset is any genre that the song's artist is associated with. Since some artists create music across multiple genres, the "genre" variable is imprecise.