Predicting the popularity of a song based on Spotify sound/musical metrics

Asen Lee

2021-12-08

Introduction

Hello, there. We're music lovers in the 21st century, so vinyls are long gone, and so are CDs. Actually, so are Walkmans and iPods. We have streaming platforms like Spotify now. You probably use either that, or Apple Music. Let's assume (like the statisticians we are) that you use Spotify. When you're listening to music on Spotify, you're probably not thinking about how it classifies each song based on a variety of criteria. But it does! Without rigorous music theory training, who knows what kind of information about music might even exist for us to know? A lot, actually. You see, Spotify evaluates each music based on elements like danceability (how suited a track is for dancing), energy (how intense and active a song is), valence (how cheerful a song makes you feel), among a slew of other parameters intended toward an API that isn't all that hard to use. Curious about this exact thing, we found a dataset on Kaggle (link: https://www.kaggle.com/sashankpillai/spotify-top-200-charts-20202021) that was web scraped from the Spotify Web API, and we began to learn what characteristics are associated with driving music that became widely popular by looking into the measurements Spotify determines and calculates with their million dollar algorithms. We think this is an extremely worthwhile project because it's immensely interesting to look into what goes into the making of a popular song.

Here is a dictionary of all metrics utilised in Spotify's database. All our variables are just metrics from Spotify, and our explanations of the variables are from the Spotify Web API documentation, found at the aforementioned link.

We want to find out how and the extent to which various audio features from Spotify metrics affect how popular a song gets.

The target/response variable:

Popularity: The popularity of the track. The value will be between 0 and 100, with 100 being the most popular.

Predictor variables: Danceability: Danceability describes how suitable a track is for dancing based on a combination of musical elements including tempo, rhythm stability, beat strength, and overall regularity. A value of 0.0 is least danceable and 1.0 is most danceable. Acousticness: A measure from 0.0 to 1.0 of whether the track is acoustic. Duration: The duration of the song in milliseconds (ms). Energy: Energy is a measure from 0.0 to 1.0 and represents a perceptual measure of intensity and activity. Typically, energetic tracks feel fast, loud, and noisy. Liveness: Detects the presence of an audience in the recording. Higher liveness values represent an increased probability that the track was performed live. Loudness: The overall loudness of a track in decibels (dB). Loudness values are averaged across the entire track. Values typical range between -60 and 0 db. Speechiness: Speechiness detects the presence of spoken words in a track. The more exclusively speech-like the recording (e.g. talk show, audio book, poetry), the closer to 1.0 the attribute value. Tempo: The overall estimated tempo of a track in beats per minute (BPM). In musical terminology, tempo is the speed or pace of a given piece and derives directly from the average beat duration. Valence: A measure from 0.0 to 1.0 describing the musical positiveness conveyed by a track. Tracks with high valence sound

more positive (e.g. happy, cheerful, euphoric), while tracks with low valence sound more negative (e.g. sad, depressed, angry). Chord:A chord is any harmonic set of pitches/frequencies consisting of multiple notes that are heard as if sounding simultaneously.

We excluded certain columns (Index, Song Name, Song ID, Release Date) available in the dataset due to them being identifiers which are meaningless for predicting. We also excluded these variables (Highest Charting Position, Number of Times Charted, Week of Highest Charting, Streams, Weeks Charted), because these variables may not be accessible or even exist for the new tracks. The rest of the excluded variables (Artist, Artist Followers, Genre) are out of scope.

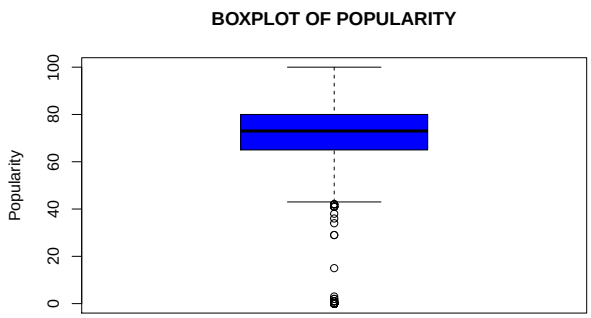
Codes and Analysis

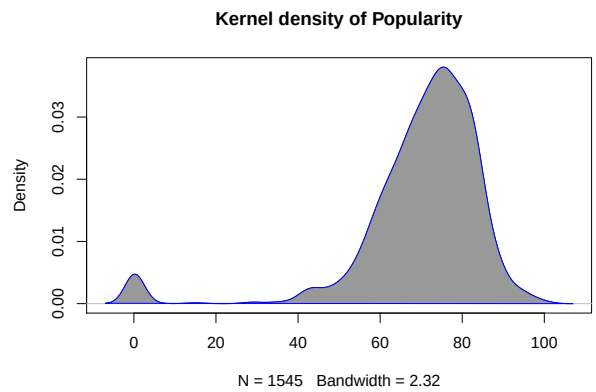
Univariate analysis:

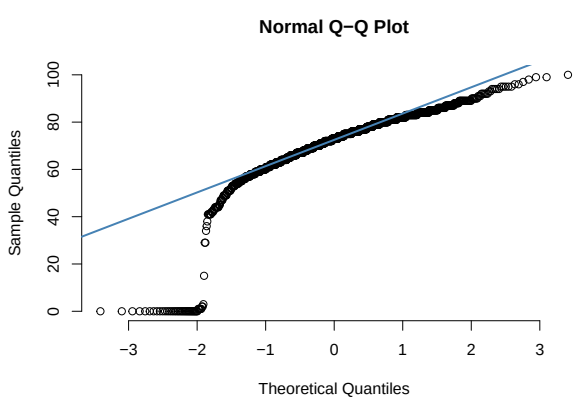
Popularity

Density plot, Box plot, and Normal Q-Q plot

```
##    vars    n  mean    sd median trimmed   mad min max range  skew kurtosis  se
## X1    1 1545 70.09 15.82     73   72.21 10.38   0 100   100 -2.46      8.5 0.4
```





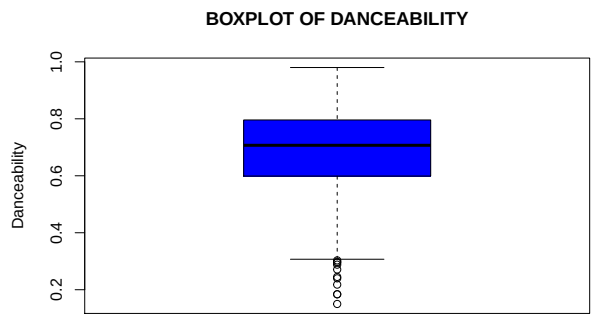


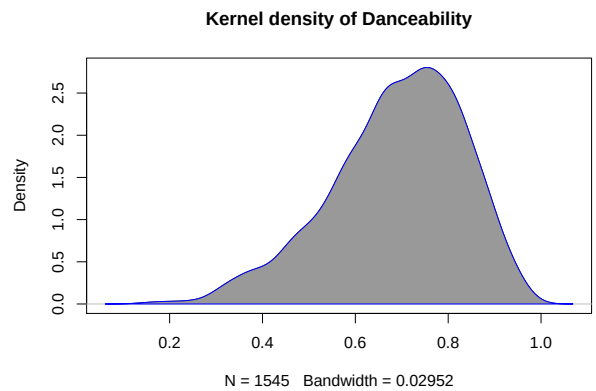
Comments on Popularity: The average popularity for the 1545 instances of data is 70.09, the median is 73.00, and the standard deviation is 15.82. Some songs are really unpopular, but most of them are between 60 and 90, so the data for popularity is extremely left skewed, as can be seen in the kernel density plot. The boxplot contains several outliers under the box as well. The QQ plot shows extreme skewness to the left and hence is not normally distributed.

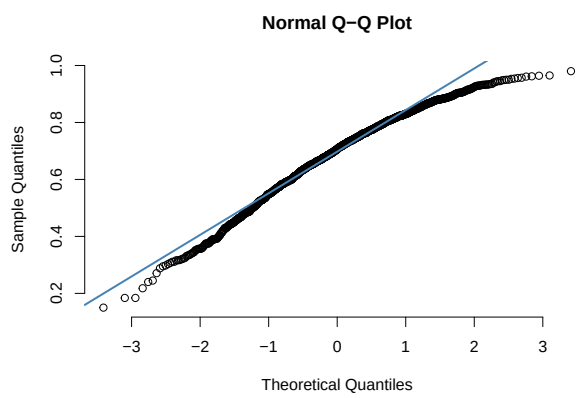
Danceability

Density plot, Box plot, and Normal Q-Q plot

```
##    vars    n mean   sd median trimmed  mad  min  max range skew kurtosis se
## X1    1 1545 0.69 0.14   0.71     0.7 0.14 0.15 0.98  0.83 -0.6      0.1  0
```





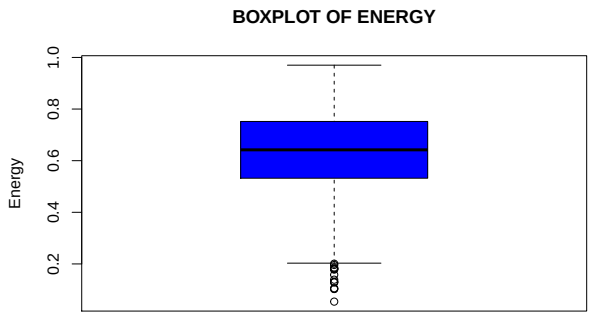


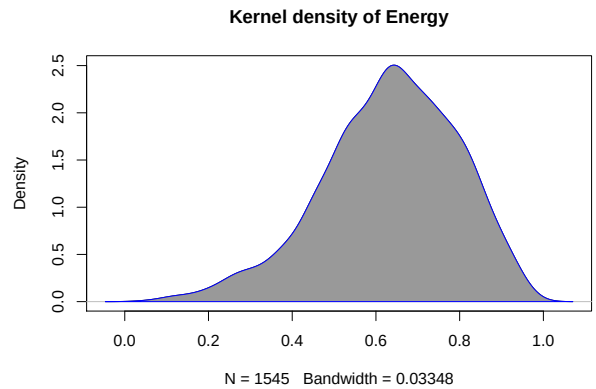
Comments on Danceability: The average danceability for the 1545 instances of data is 0.69, the median is 0.71, and the standard deviation is 0.14. The data for danceability is slightly left skewed, as can be seen in the kernel density plot. The boxplot contains some outliers under the box as well. The QQ plot shows very slight skewness to the left, but is almost normally distributed.

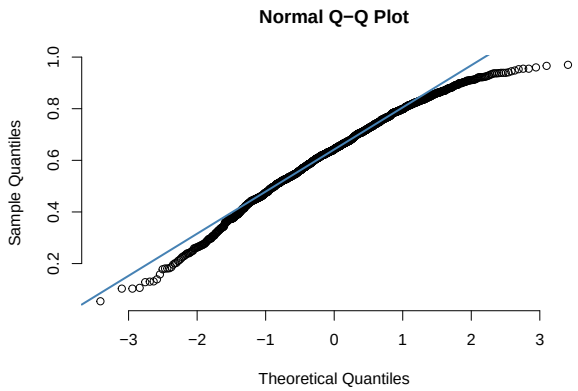
Energy

Density plot, Box plot, and Normal Q-Q plot

```
##    vars    n mean   sd median trimmed  mad  min  max range  skew kurtosis se
## X1    1 1545 0.63 0.16   0.64    0.64 0.16 0.05 0.97  0.92 -0.48     0.08  0
```

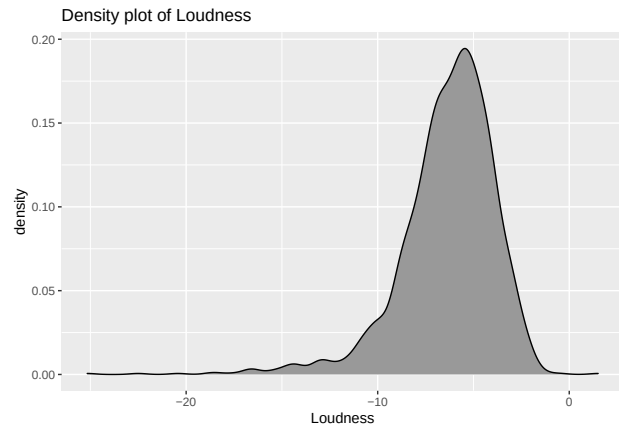




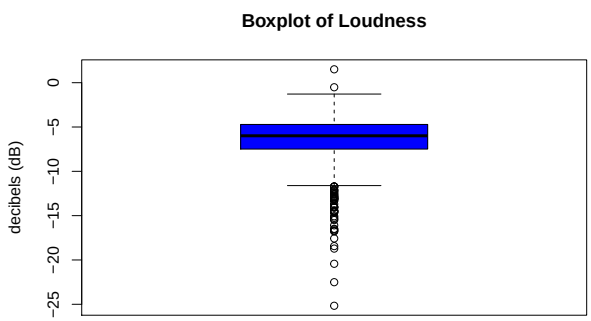


Comments on Energy: The average energy for the 1545 instances of data is 0.63, the median is 0.64, and the standard deviation is 0.16. The data for energy is only slightly left skewed, as can be seen in the kernel density plot. The boxplot contains some outliers under the box as well. The QQ plot shows very slight skewness to the left, but is almost normally distributed.

Loudness

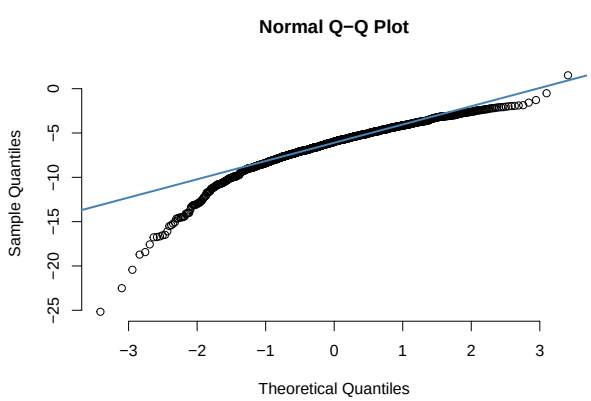


Density plot, Box plot, and Normal Q-Q plot



Comments on Loudness: From the density plot above, we can see that "Loudness" has a mean around -6.35 and is skewed to the left. Most of the observations are between the range of -10 to -2.5 dB.

A similar trend can be found from the boxplot with the mean being around -6.35 and the plot showing a left skewed distribution.



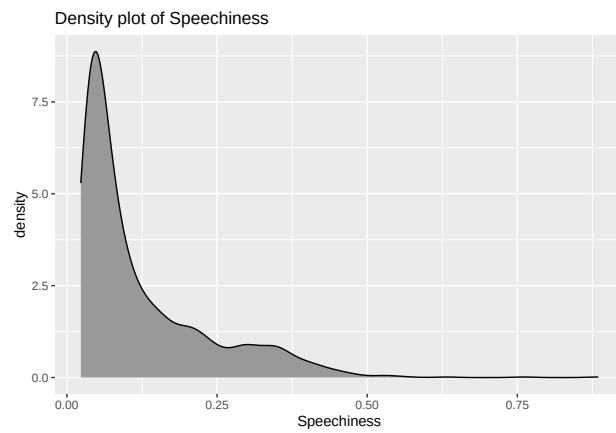
Comments on Q-Q plot: The QQ plot shows skewness to the left and hence is not normally distributed.

Statistic	Value
Mean	-6.348

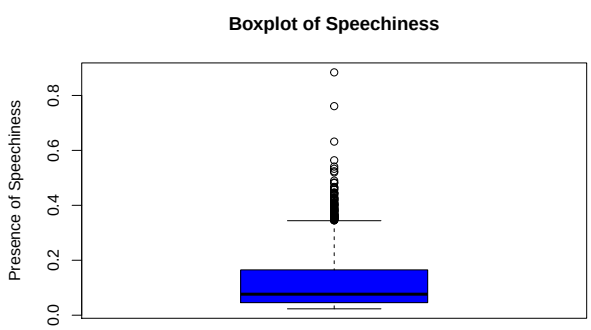
Statistic	Value
Median	-5.99
STD	2.509
Skewness	-1.571
SE	0.064

Comments on Loudness: The predictor, "Loudness", is measured in decibels (dB), and it has a mean of -6.35dB and a median of -5.99dB. The minimum value is found to be -25.17dB and the maximum value is 1.51dB. The feature has a standard deviation of 2.51 and a standard error of 0.064.

Speechiness

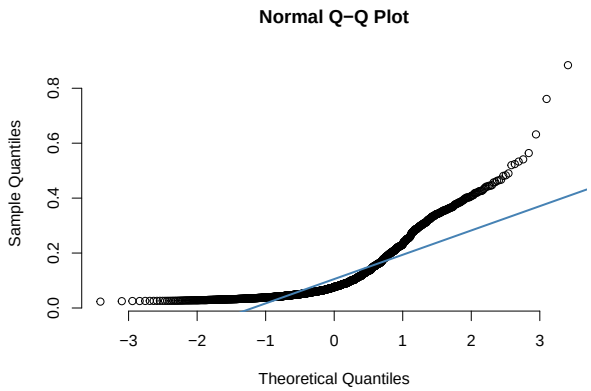


Density plot, Box plot, and Normal Q-Q plot



Comments on Speechiness: From the density plot above, we can see that "Speechiness" has a mean around 0.1 and is skewed to the right Most of the observations are between the range of 0.0 to 0.35.

A similar trend can be found from the boxplot with the mean being around 0.1 and the plot showing a right skewed distribution.

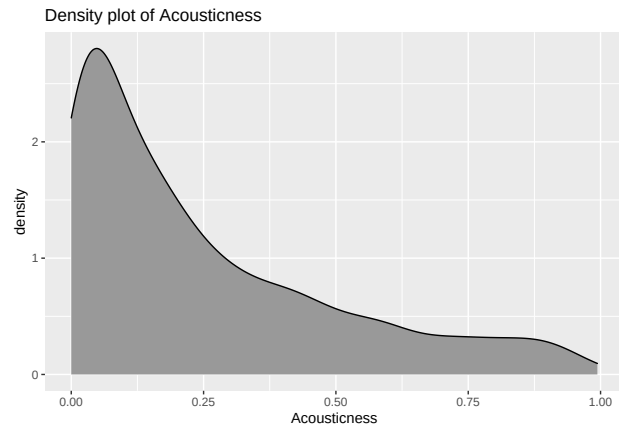


Comments on Q-Q plot: The QQ plot show skewness to the right and hence is not normally distributed.

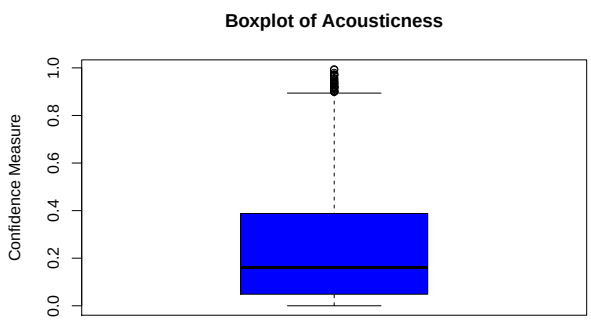
Statistic	Value
Mean	0.124
Median	0.076
STD	0.11
Skewness	1.674
SE	0.003

Comments on Speechiness: The predictor, "Speechiness", is measured by the presence of spoken words in a track. For example, a value close to 1.0 indicates speech-like recordings such as talk shows and values between 0.66 to 0.33 may indicate rap music. Values lower than 0.33 are music and non-speech-like tracks. The predictor has a mean of 0.12 and a median of 0.08. The minimum value is found to be 0.02 and the maximum value is 0.88. The feature has a standard deviation of 0.11 and a standard error of 0003.

Acousticness

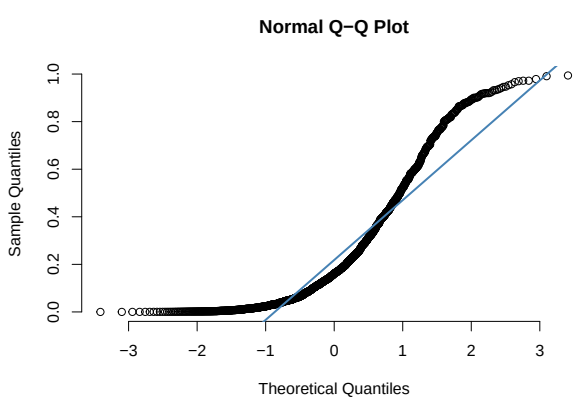


Density plot, Box plot, and Normal Q-Q plot



Comments on Acousticness: From the density plot above, we can see that "Acousticness" has a mean around 0.25 and is skewed to the right. Most of the observations are between the range of 0.0 to 0.4.

A similar trend can be found from the boxplot with the mean being around 0.25 and the plot showing a right skewed distribution.



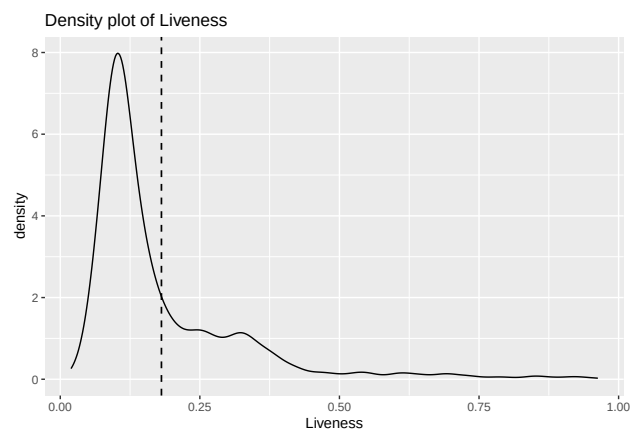
Comments on Q-Q plot: The QQ plot shows skewness to the right and hence is not normally distributed.

Statistic	Value
Mean	0.249

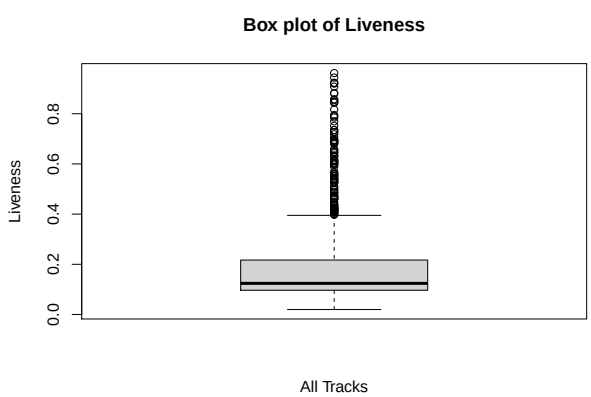
Statistic	Value
Median	0.161
STD	0.25
Skewness	1.148
SE	0.006

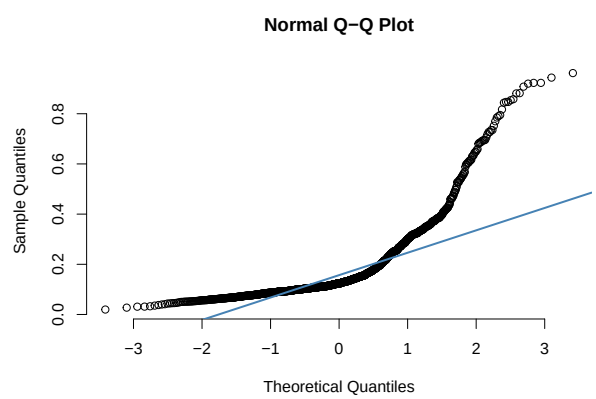
Comments on Acousticness: The predictor, "Acousticness", is measured by a confidence measure from 0.0 to 1,0 of whether the track is acoustic or not. It has a mean of 0.25 and a median of 0.16. The minimum value is found to be 0.0 and the maximum value is 0.99. The feature has a standard deviation of 0.25 and a standard error of 0.006.

Liveness



Density plot, Box plot, and Normal Q-Q plot

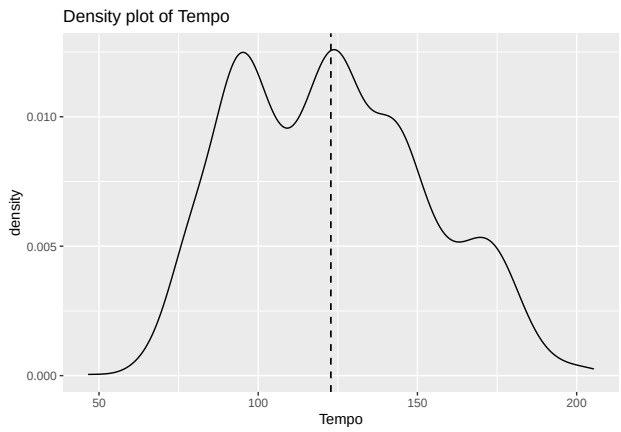




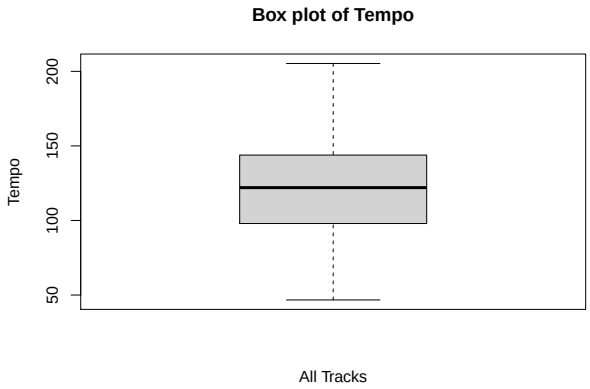
Statistic	Value
Mean	0.181
Median	0.124
STD	0.144
Skewness	2.409

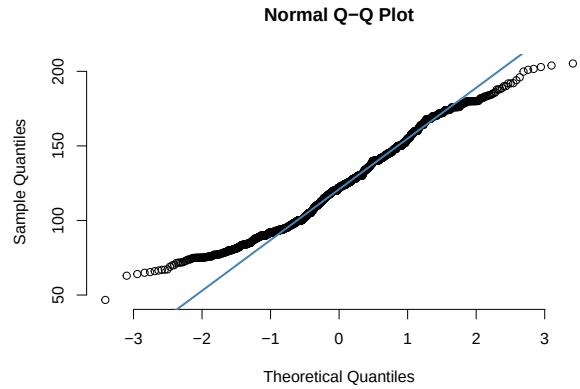
Comments on Liveness: Liveness is a measure of the probability of live performance. The distribution is very right-skewed and deviated from standard normal distribution. This indicates that more tracks in the chart are considered "studio".

Tempo



Density plot, Box plot, and Normal Q-Q plot

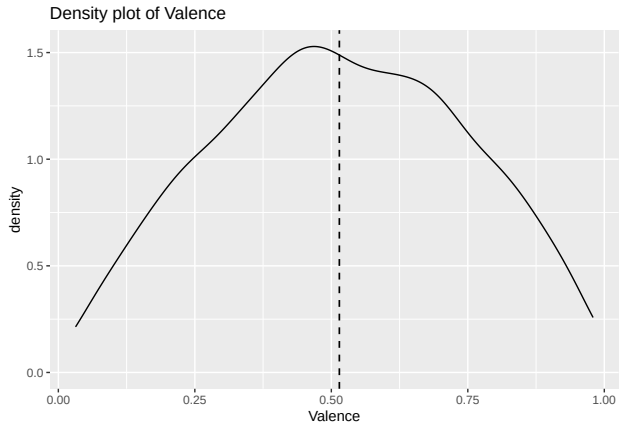




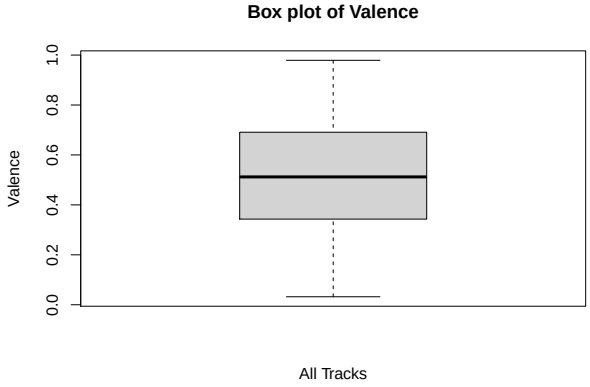
Statistic	Value
Mean	122.811
Median	122.012
STD	29.591
Skewness	0.294

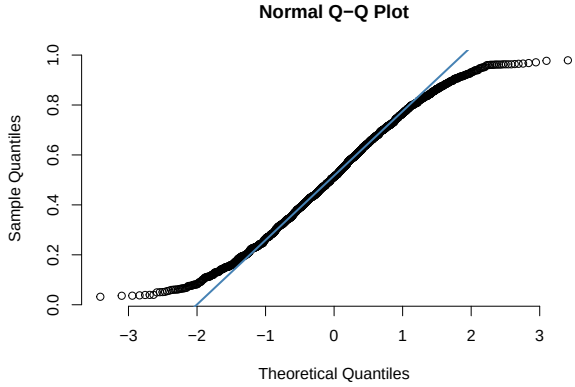
Comments on Tempo: Tempo is the average speed of a song measured in bpm. The skewness is only 0.3 which can be considered symmetrical. The shape of the distribution is multimodal (2-3 peaks). The distribution of the data mildly deviated from standard normal distribution. Based on the tempo markings, The mean tempo (122) lies in the Allegro (Fast) marking (120-168bpm).

Valence



Density plot, Box plot, and Normal Q-Q plot

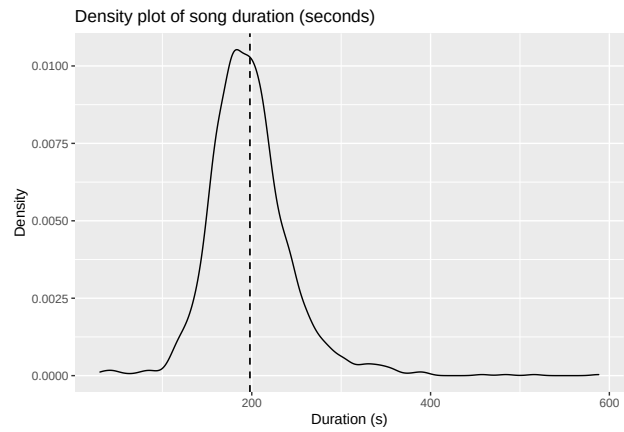




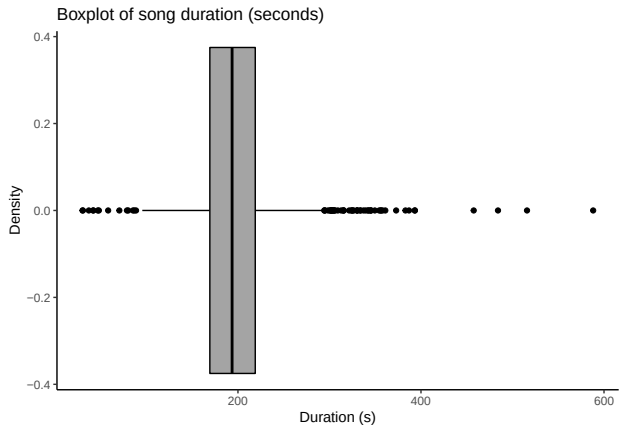
Statistic	Value
Mean	0.515
Median	0.512
STD	0.227
Skewness	-0.025

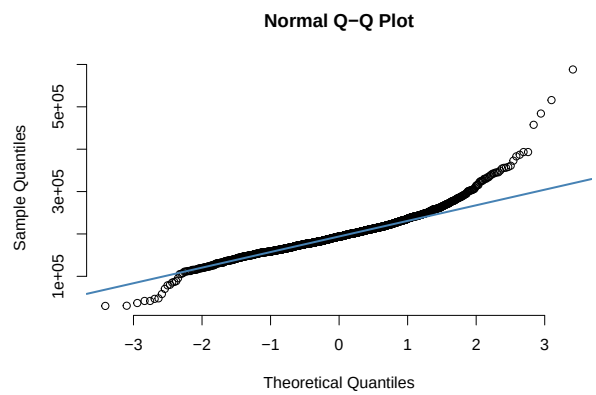
Comments on Valence: Valence measures the positiveness of the track. The skewness and shape shows the data distribution of this variable is very symmetrical. Overall the majority of the data seems aligned with standard normal distribution. And the line of symmetry locates at the middle (valence = 0.5) which means the mean positiveness of the charted track is very neutral.

Song Duration



Density plot, Box plot, and Normal Q-Q plot



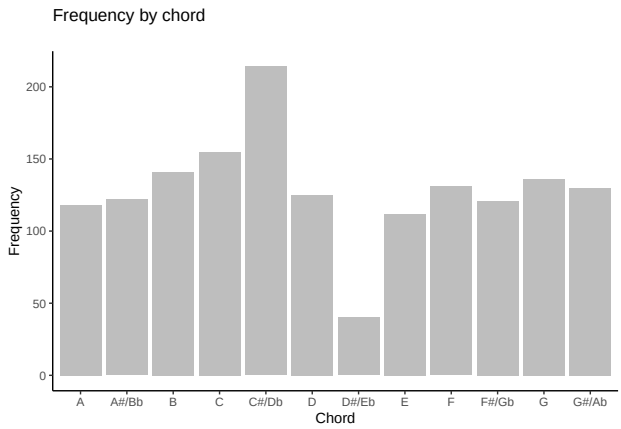


Some summary statistics all in seconds

Statistic	Value (s)
Mean	197.941
SD	47.149
Min	30.133
1st Quartile	169.266
Median	193.591
3rd Quartile	218.902
Max	588.139

Comments on Song Duration: The mean song duration is 198 seconds with a standard deviation of 47 seconds. The song duration is heavily concentrated between 100 and 300 seconds, and it's clear the data is not normally distributed as further evidenced by the heavy tails in our qq plot. There is far less data at the tails of this distribution then in a normal distrbution. Moreover, the data is right skewed which is expected as songs are generally short and have a minimum possible duration (0 seconds), but not a maximum.

Chord



Frequency plot

Some summary statistics

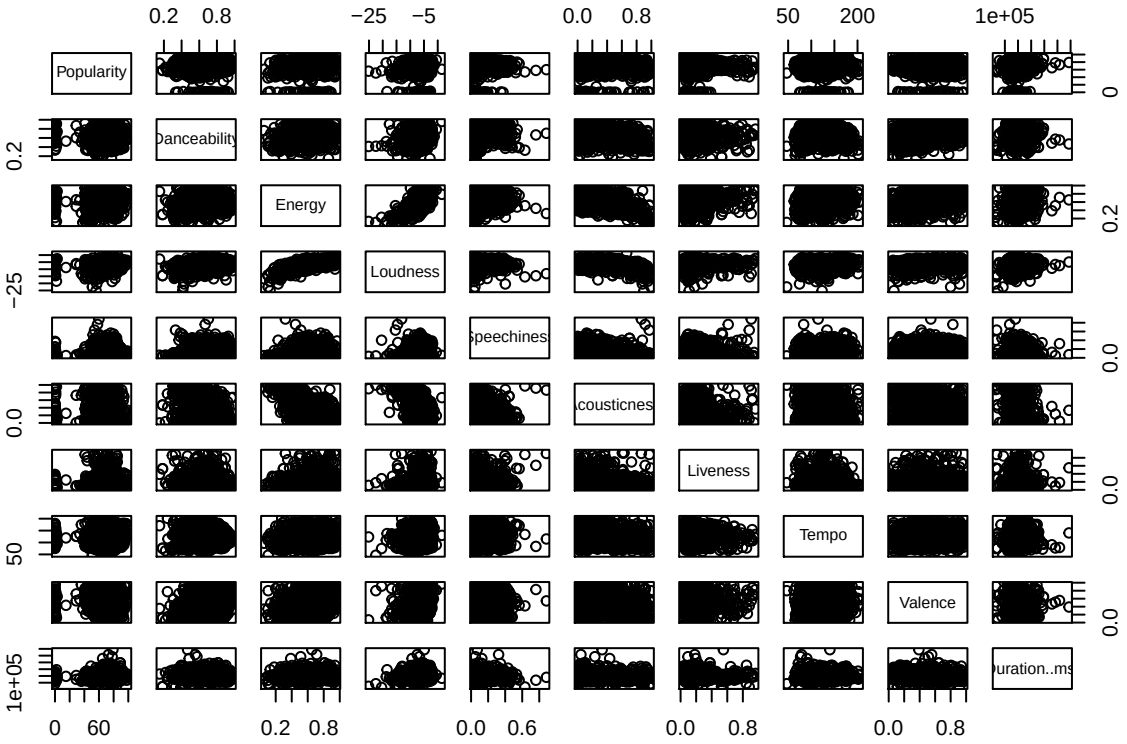
Chord	Counts	Percentages (%)
A#/Bb	122	8.549
B	141	9.881
C	155	10.862
C#/Db	214	14.996
D	125	8.76
D#/Eb	40	2.803
E	112	7.849
F	131	9.18
F#/Gb	121	8.479
G	136	9.53
G#/Ab	130	9.11

Comments on Chord: The majority of chords comprise around 8% of all songs. Interestingly, there are significantly more songs in the C#/Db chord (14%), and significantly less songs in the D#/Eb chord (3%).

Bivariate analysis:

Pairwise Correlation Table and Scatter Plots Since Chord is a categorical variable, we will do one analysis excluding chord. Then another, including it.

	Pop.	Dan.	Ene.	Lou.	Spe.	Aco.	Liv.	Tem.	Val.	Dur.
Popularity	1.00	0.03	0.09	0.16	-0.03	-0.09	-0.03	-0.02	0.00	0.08
Danceability	0.03	1.00	0.14	0.23	0.24	-0.32	-0.11	-0.04	0.36	-0.10
Energy	0.09	0.14	1.00	0.73	0.02	-0.54	0.12	0.11	0.36	0.06
Loudness	0.16	0.23	0.73	1.00	-0.02	-0.48	0.04	0.10	0.30	0.08
Speechiness	-0.03	0.24	0.02	-0.02	1.00	-0.13	0.07	0.11	0.04	-0.09
Acousticness	0.09	-0.32	-0.54	-0.48	-0.13	1.00	-0.01	-0.06	-0.10	-0.05
Liveness	-0.03	-0.11	0.12	0.04	0.07	-0.01	1.00	-0.02	0.01	0.02
Tempo	-0.02	-0.04	0.11	0.10	0.11	-0.06	-0.02	1.00	0.06	0.00
Valence	0.00	0.36	0.36	0.30	0.04	-0.10	0.01	0.06	1.00	-0.12
Duration..ms	0.08	-0.10	0.06	0.08	-0.09	-0.05	0.02	0.00	-0.12	1.00

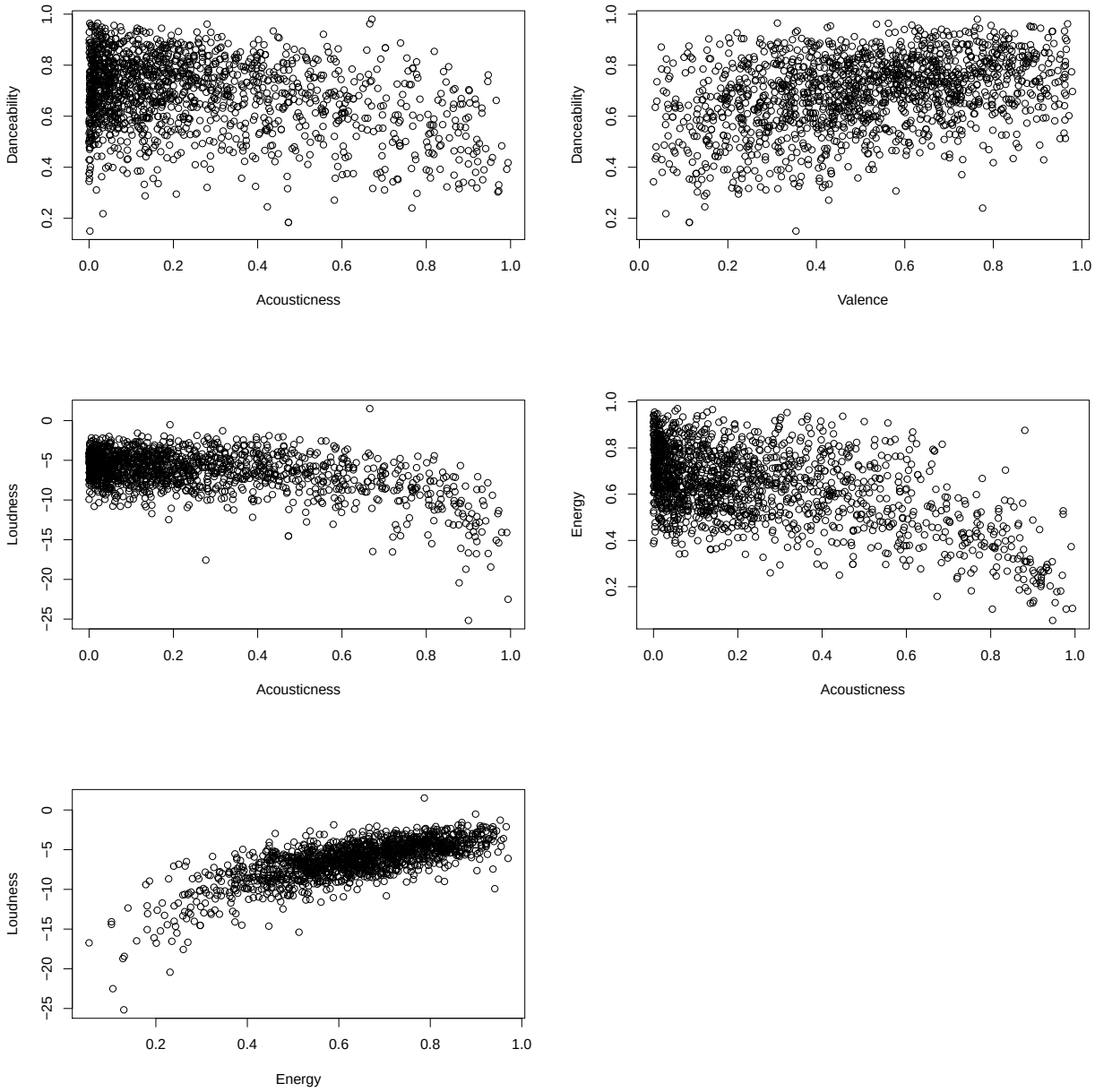


Comments: Most of the data distributions are rather random with no specific trend. This indicates that Simple Linear Regression may not be a good candidate model.

We'll consider variables with a magnitude of correlation between 0.3 and 0.5 to have low correlation, 0.5 and 0.7 to have moderate correlation, and 0.7 and 0.9 to have high correlation.

Low correlation: Acousticness and danceability (-0.32). Danceability and valence (0.36). Acousticness and loudness (-0.48).
Moderate correlation: Acousticness and energy (-0.54).
High correlation: Energy and loudness (0.73).

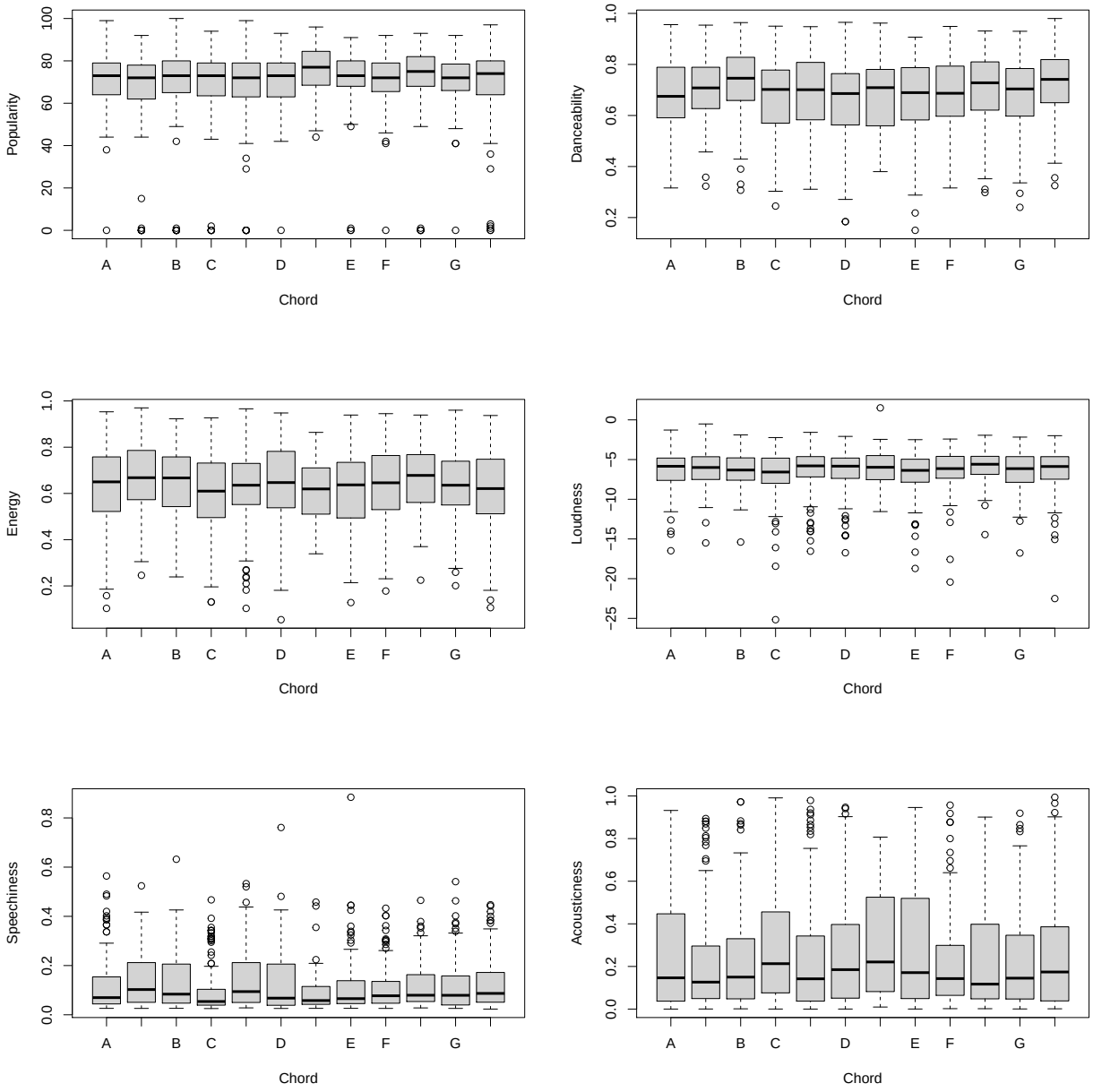
Plots of most correlated variables

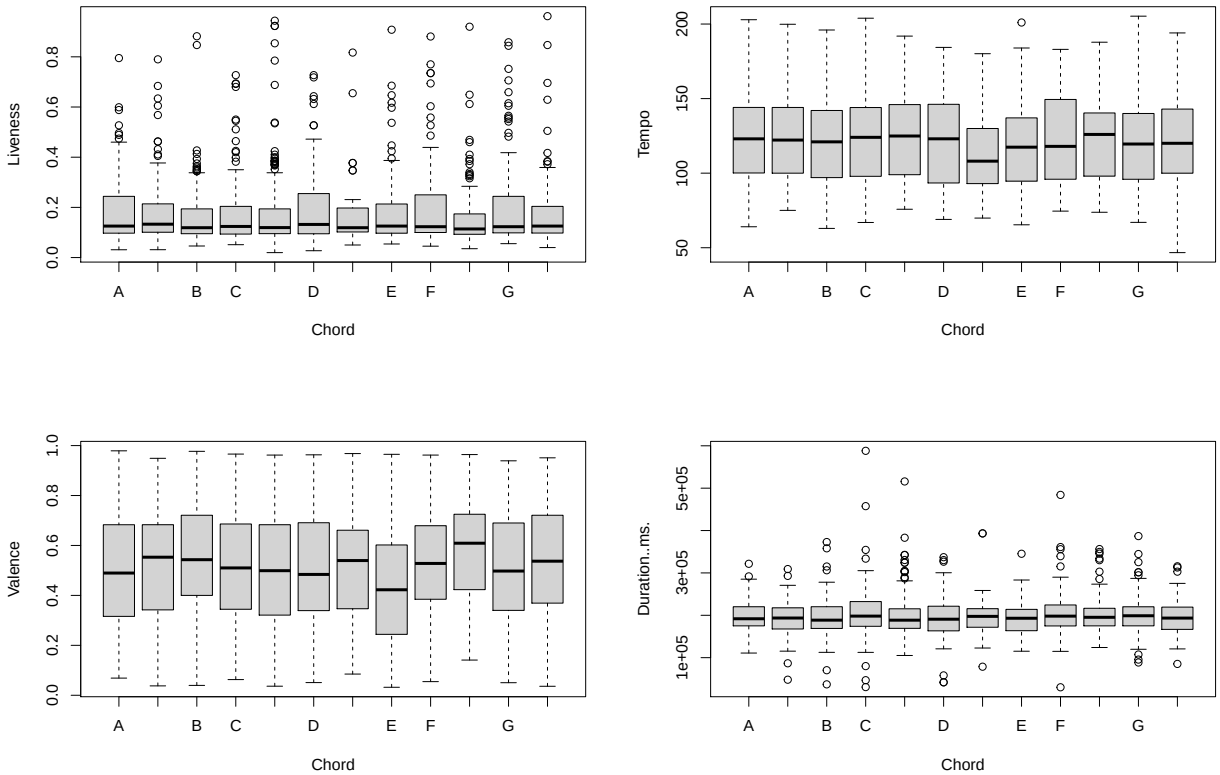


Comments on Correlated Variables: These correlations make sense. Songs that have one of the features among danceability, enery, loudness and valence will tend to exhibit the other features from anecdotal experience. By similar token, if the song is more acoustic, it will likely be less energetic and loud explaining the negative correlation.

Now, we examine how each variable is distributed by its chord.

Boxplots of metrics by chord





Comments on Metrics by Chord: There is no apparent trend or notable observation for the distributions of the variables separated by its chord. For the purposes of our research question, we could point out that songs in D#/Eb tend to have slightly more popularity.