Exploratory Data Analysis Final Report

Musical Genre Trends Across The Ages!

12-07-2021

Anna Cameron, Enya Tan, Mubarak Ganiyu, Sriram Kannan

Introduction:

Music serves as an art form and a cultural activity. Under its category there are many far-reaching influences, appealing features, and influential musical art forms. Music has been evolving and changing since the beginning of human civilization. With the continuous development of society, people's spiritual pursuit is constantly strengthened while music also flourishes, which is deeply rooted in our lives. This kind of artistic performance is one of the ways people use to edify their sentiment, relax, and entertain themselves.

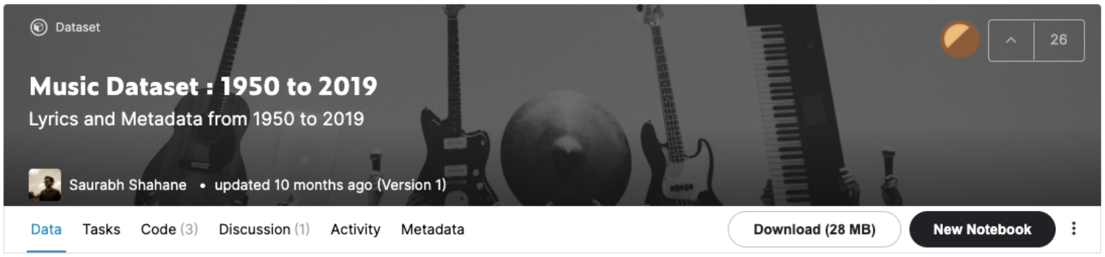
This project uses exploratory data analysis methods to find out the changes in music genre trends from 1950 to 2019, including exploration of the most produced genre of music, correlational analysis of music features across genres, the difference across genres on average based on different features, the changes of genres on average over time about different musical features, the variation of genres with regards to the songs' contents, difference between genres in terms of the topic, and so on. Different types of visuals are created to present the analysis topics and solve the analysis problems.

In developing this project, we first established the project topic and collected relevant data. Then we used R software to carry out steps of data cleaning, variable transformation, and creation of new variables. Next we used R software to conduct exploratory data analysis of our analytical problems to find useful insights. By analyzing this topic, people can better understand the changes of music genre trends from the past to today, and it may also serve as a reference for the development of music in the future.

Data:

Our data set was sourced from *Kaggle* which is a collaborative website used to publish data sets and machine learning practices among data scientists. The data set was created by Sharabh Shahane and includes information from songs between 1950 and 2019.

The data set contains track names, artists, release year, song length, genre, and even the lyrics of the songs. The data set also contains variables like sadness, feelings, communication, and romance which relate to each song's subject matter. The songs included in the data set were also given scores between 0 and 1 based on the danceability, acousticness, valence, instrumentalness, energy, and loudness.



Dataset

Music Dataset : 1950 to 2019

Lyrics and Metadata from 1950 to 2019

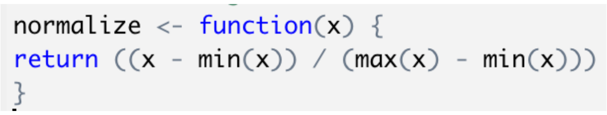
Saurabh Shahane • updated 10 months ago (Version 1)

Term	Definition
Danceability	Suitability 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 is least danceable and 1 is most danceable.
Acousticness	An estimate of how acoustic a particular song is. Songs with high 'acousticness' will consist mostly of natural acoustic sounds (think acoustic guitar, piano, orchestra, the unprocessed human voice), while songs with a low 'acousticness' will consists of mostly electric sounds (think electric guitars, synthesizers, drum machines, auto-tuned vocals and so on).
Valence	Describes the musical positiveness conveyed by a track. Tracks with high valence (closer to 1) sound more positive (e.g. happy, cheerful, euphoric), while tracks with low valence (closer to 0) sound more negative (e.g. sad, depressed, angry).
Instrumentalness	Predicts whether a track contains no vocals. "Ooh" and "aah" sounds are treated as instrumental in this context. Rap or spoken word tracks are clearly "vocal." The closer the instrumentalness value is to 1 the greater likelihood the track contains no vocal content. Values above 0.5 are intended to represent instrumental tracks, but confidence is higher as the value approaches 1.
Energy	Energy is a measure from 0 to 1.and represents a perceptual measure of intensity and activity. Typically, energetic tracks feel fast, loud, and noisy.
Loudness	Relative amplitude of the song in decibels (DB) normalized to values between 0 and 1.

Data Cleaning and Transformation:

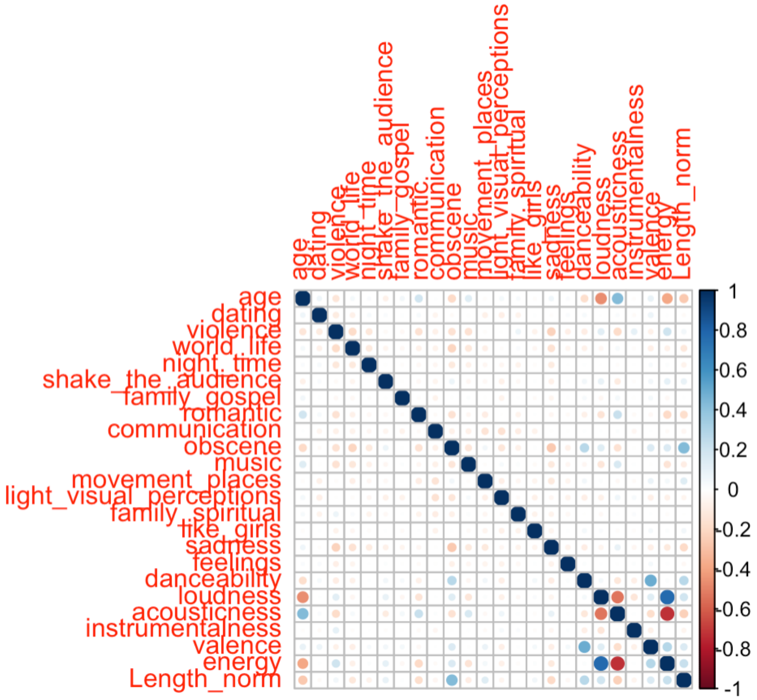
The original version of our music data set included 28,372 observations across 31 different variables. The first step we took to prepare our data for analysis was to use *Assertr* to check for any missing information. Some of the variables in our data set like danceability, acousticness, and valence were values that needed to fall between 0 and 1, so we also used *Assertr* to verify that the recorded values fell within that range.

Next, we created a normalizing function to restrict values of variables between 0 and 1 using the Min-Max Scaler. This was done to enable correlation analysis which included these variables. Using the normalize function, we created a new variable called "Length_norm" which represented the normalized length for each individual song.



```
normalize <- function(x) {
return ((x - min(x)) / (max(x) - min(x)))
}
```

We also added a variable with new ID numbers for each track, and we created a new variable called "Age_actual". "Age_actual" was calculated by subtracting the release year of the song from 2019. While the Original Age variable was already in a scaled format, it could be used directly in correlation analysis but could not be interpreted and hence, "Age_actual" was created to solve this.



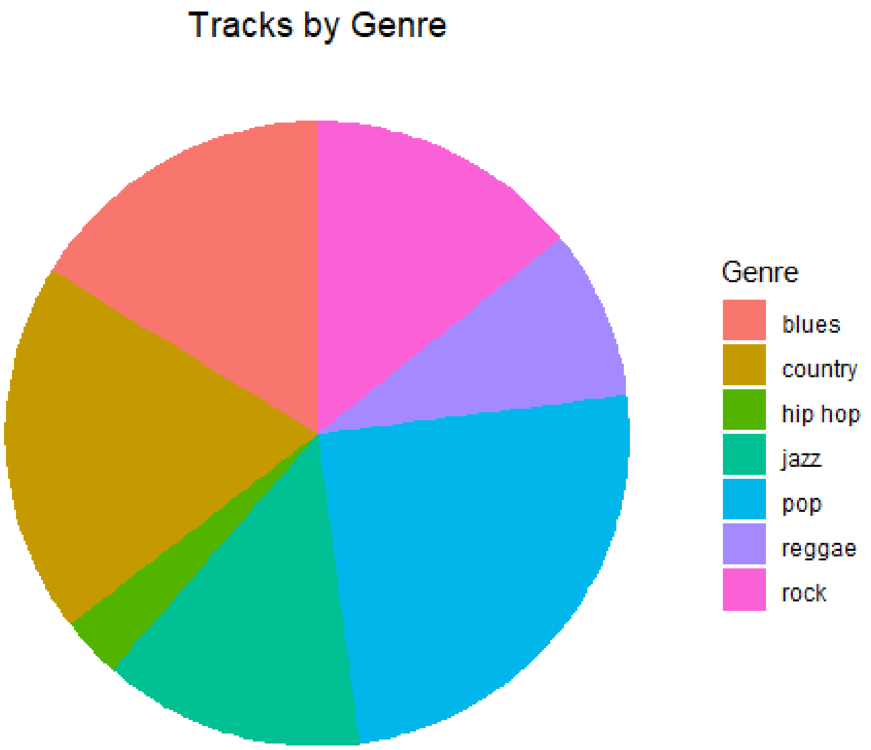
The final step of our data cleaning process involved checking the correlations between the different variables in the original data set. This allowed us to determine which variables should be removed from the data set prior to performing our analysis. From our correlation coefficient matrix we decided to remove 16 unnecessary or redundant variables. We found that the strongest relationships were among danceability, acousticness, loudness, valence, energy, age and instrumentalness, so we decided to make those variables the main focus of our analysis. With the addition of our three new variables (Age_actual, Length_norm, and ID) and the removal of unnecessary variables, our final data set had 18 variables as opposed to the original 31 variables.

Analysis:

In order to conduct a holistic analysis on musical genres across the ages, numerous forms of analysis were conducted with the objective of figuring out how the genres have evolved over time, how different variables relate to one another by genre and how genres differ according to different parameters. Important questions were laid out in order to address the different objectives stated above.

Which Genre of music has been produced the most?

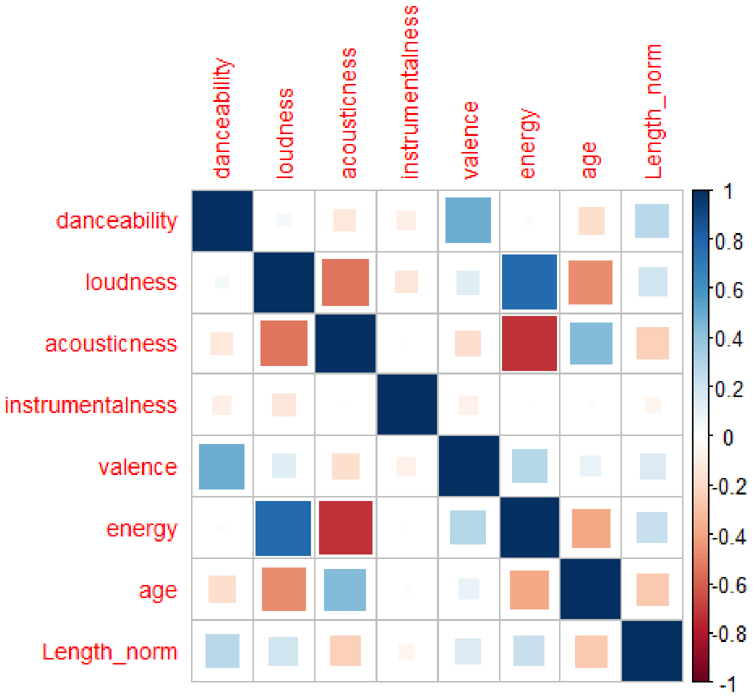
The analysis began by figuring out how many songs each genre has produced. The purpose of this analysis was to have a broad understanding of how diverse the data set was with regards to genre. Looking at the pie chart below, most of the music released belongs to the pop genre. Pop music can be categorized as popular music. It is a form of music that incorporates different styles of music. Due to its loose definition, it is no surprise that a lot of music is categorized as pop. Hip hop had the least share of the pie. This could be due to the fact that hip hop is a relatively new genre that sprung up in the mid 90s. Hence, not that many songs that have been released belong to this genre.



Correlational Analysis of Musical Features across Genres

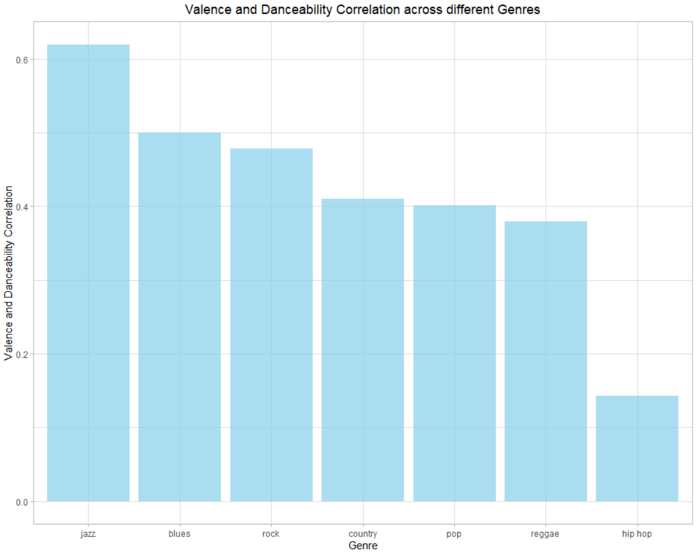
(1) Correlation Coefficient Matrix

The correlation coefficient matrix figure below shows the correlation coefficient relationship between every pair of musical feature variables. It can be seen from the figure that there exists a strong positive correlation between some variables, such as loudness and energy; there is also a strong negative correlation between some variables, such as energy and acousticness. In addition, there are some variables that do not have obvious correlation with others. This figure provides a reference for our subsequent analysis.



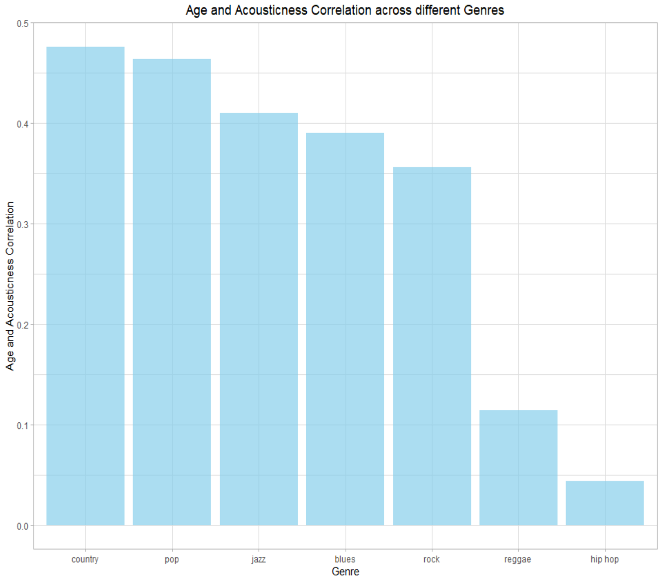
(2) Valence and Danceability:

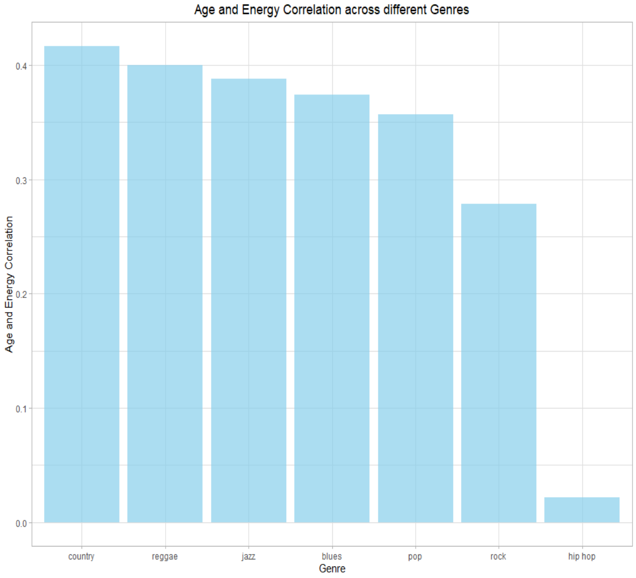
Valence is a measure of how positive a song is, and danceability is a measure of how likely people are to dance to it. Reviewing the valence-danceability correlations across genres, jazz contributes to this correlation the most with a correlation of over 0.6. This means that the jazz songs that are positive are easy to dance and vibe to. However, this is not the case for hip hop as it ranks last among the genres in this category because its valence-danceability correlation is very weak. Hence, for hip hop songs, there might not be a relationship between valence and danceability. Blues, rock, country, pop and reggae sit in second, third, fourth, fifth and sixth place respectively for this category.

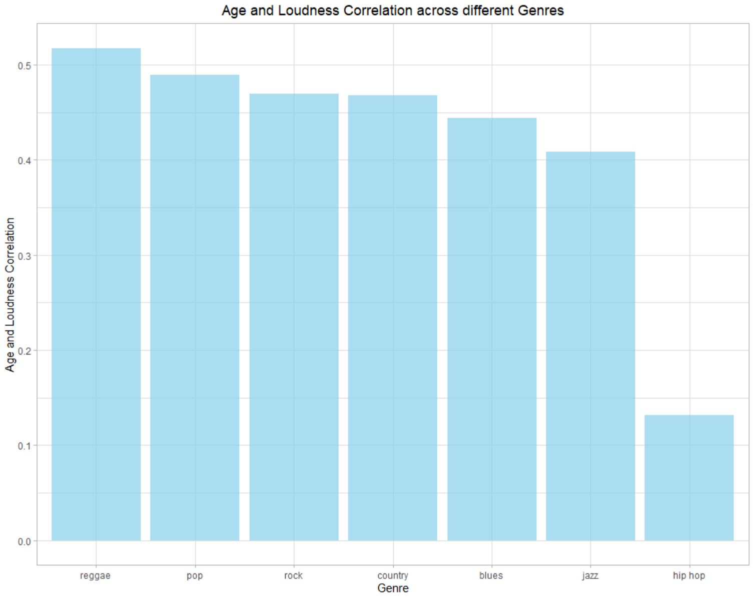


(3) Age and Feature(s) Correlations:

Similarly, the figures below show the correlation between age and acousticness across different genres, age and energy correlation across different genres and age and loudness correlation across different genres. Country music makes the greatest contribution to the correlation between age and acousticness, age and energy. Reggae contributes most to the correlation between age and loudness while its contribution is much lower when it comes to age and acousticness. We can also see that hip hop contributes the least to all the correlation relationships between the variables provided below. Hip-hop music seems to be relatively independent in terms of those musical features, while country music is the opposite.

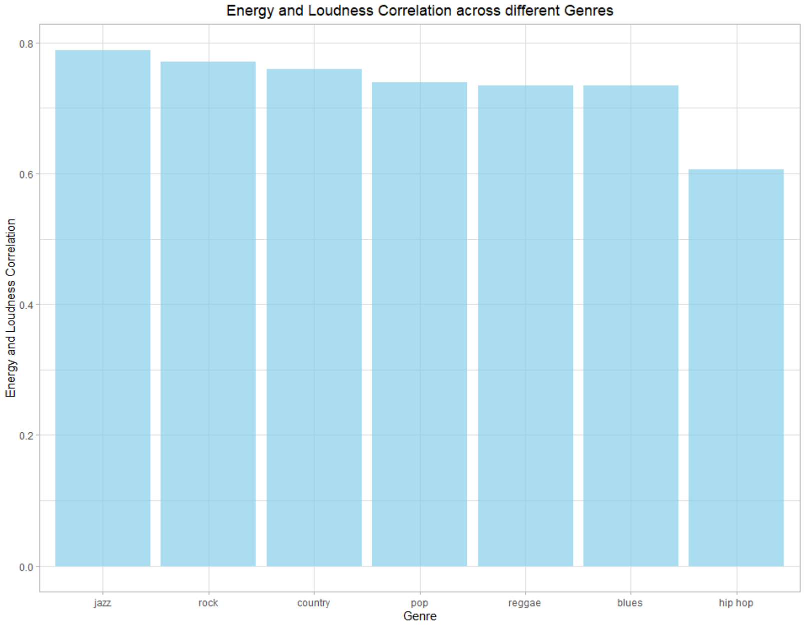






(4) Energy and Loudness Correlation:

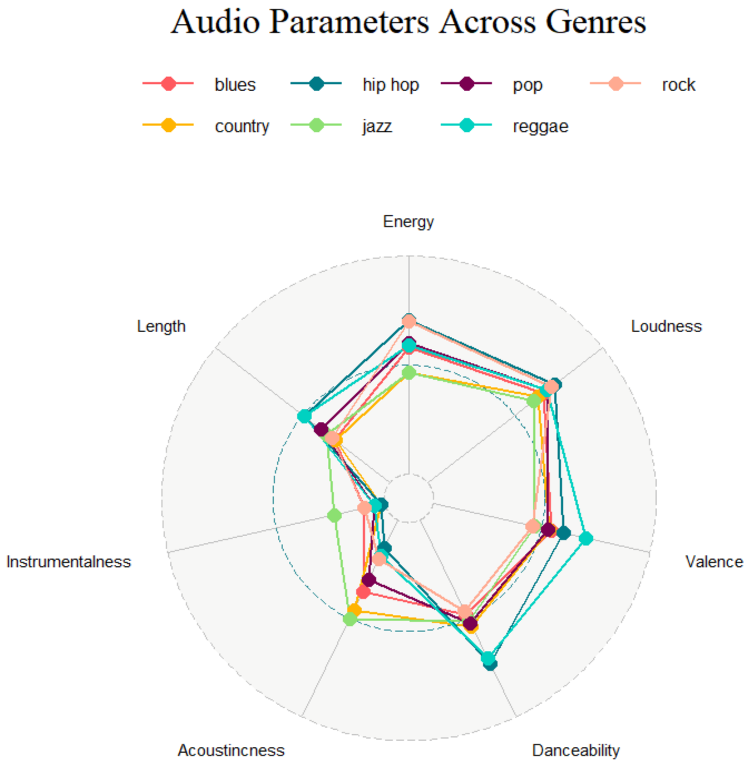
The correlation between energy and loudness was the highest upon observing the correlation coefficient matrix. This correlation was about 0.77. Like the valence-danceability correlations across genres, jazz contributes to this correlation the most with a correlation of almost 0.8. This means that songs that are produced to be lively tend to have high pitch. Rock is second in this category with a correlation of 0.77. Country and Pop are third and fourth respectively with correlations of 0.76 and 0.74 respectively. Reggae sits in fifth place with a correlation of 0.7347. The last two are blues and hip hop with correlations of 0.7338 and 0.6 respectively.

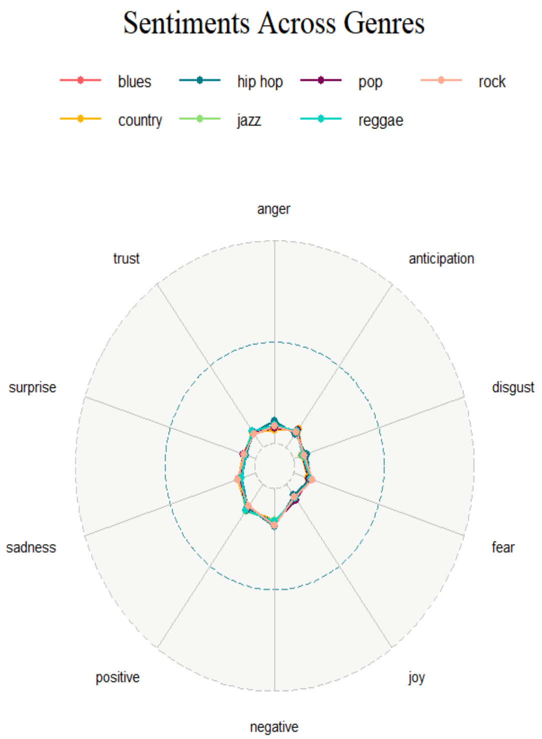


Exploring How Genres Differ Based on Different Audio Parameters

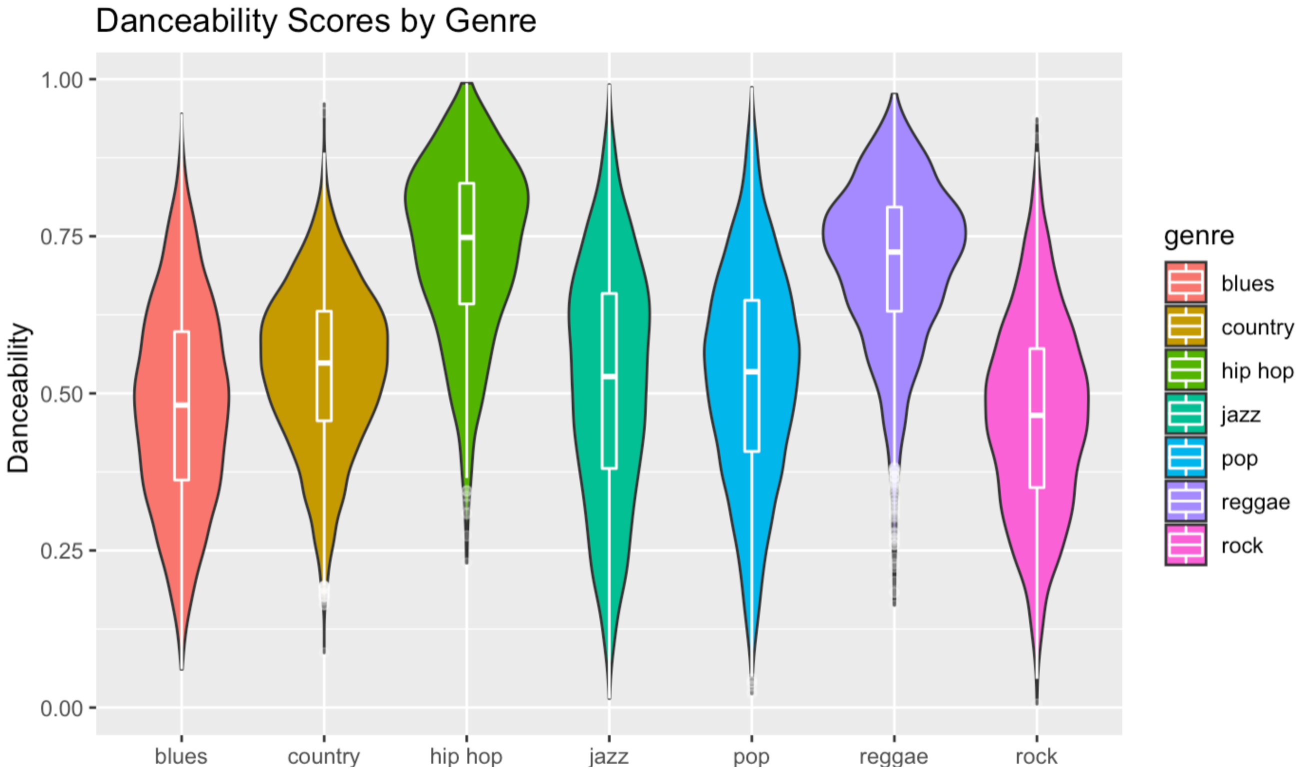
From the radar plots below, we can see how genres vary based on various audio parameters. Hip hop and reggae have the highest danceability by a significant margin among all the other genres with rock scoring the lowest; hip hop and rock have the highest energy and loudness with jazz being the lowest in both of these categories; jazz has the highest instrumentalness by a large margin with hip hop being the lowest while in the case of acousticness, it's again jazz with the highest closely followed by country while hip hop has the least. Reggae seems to have the longest songs as well. This can help us understand the characteristics of different genres of music.

In case sentiment analysis, according to the sentiment radar plot, different genres of music, overall, seem to express very similar emotional tendencies which was an interesting result as you would normally think that wouldn't be the case. The plot below was created by calculating the ratio of each sentiment within a genre as opposed to directly comparing the number of each sentiment across genres as the results would get skewed to genres with the larger number of songs in the dataset.





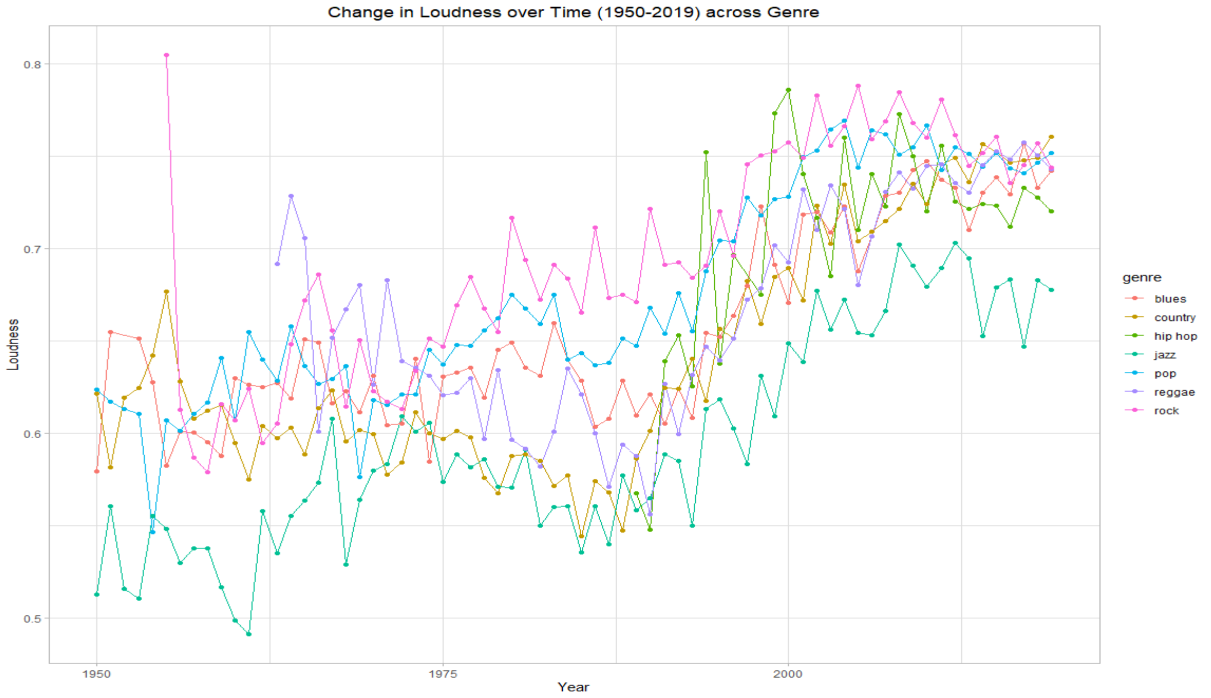
For some listeners, danceability is a really important aspect of a song, so we wanted to compare danceability across different music genres. Based on our violin plot, the average danceability scores across country, jazz, and pop appear to be pretty similar while hip hop and reggae have higher average danceability scores. However, blues, jazz, pop, and rock have a greater spread of danceability scores than country, hip hop, and reggae.



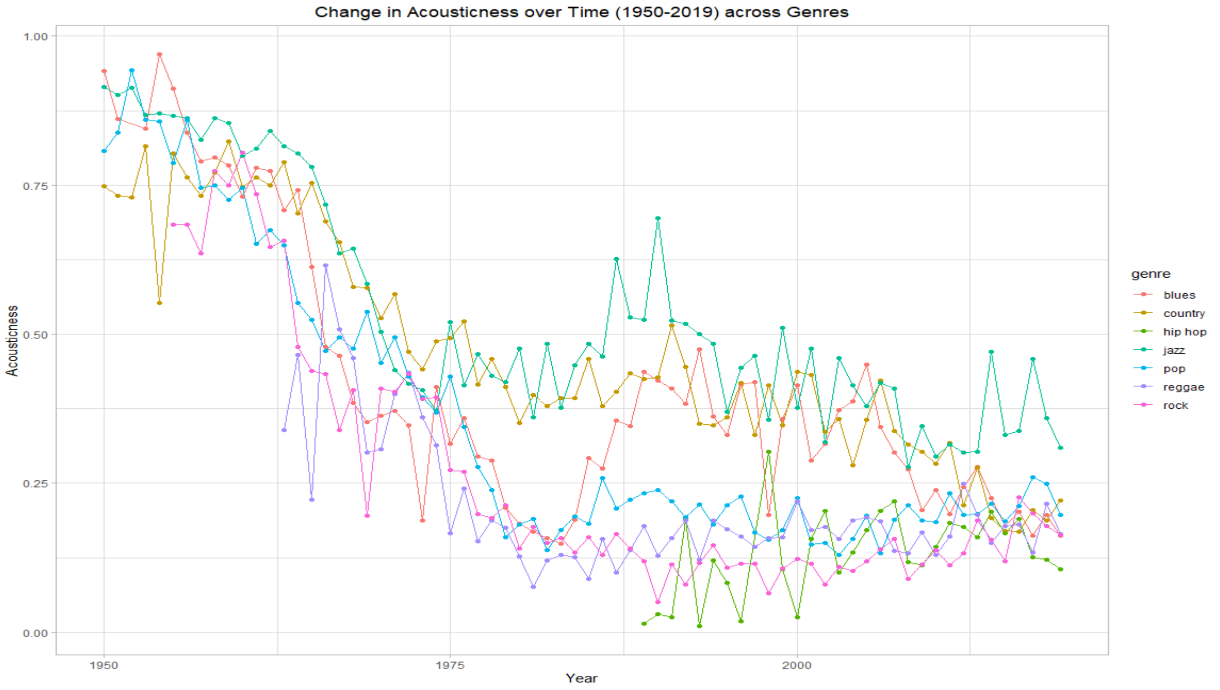
Genre Changes on Average over Time According to Different Musical Features

The top three correlated relationships from the correlation coefficient matrix were between energy and loudness, loudness and acousticness, and energy and acousticness. Hence, it was imperative to see how these three variables have changed over time across genres.

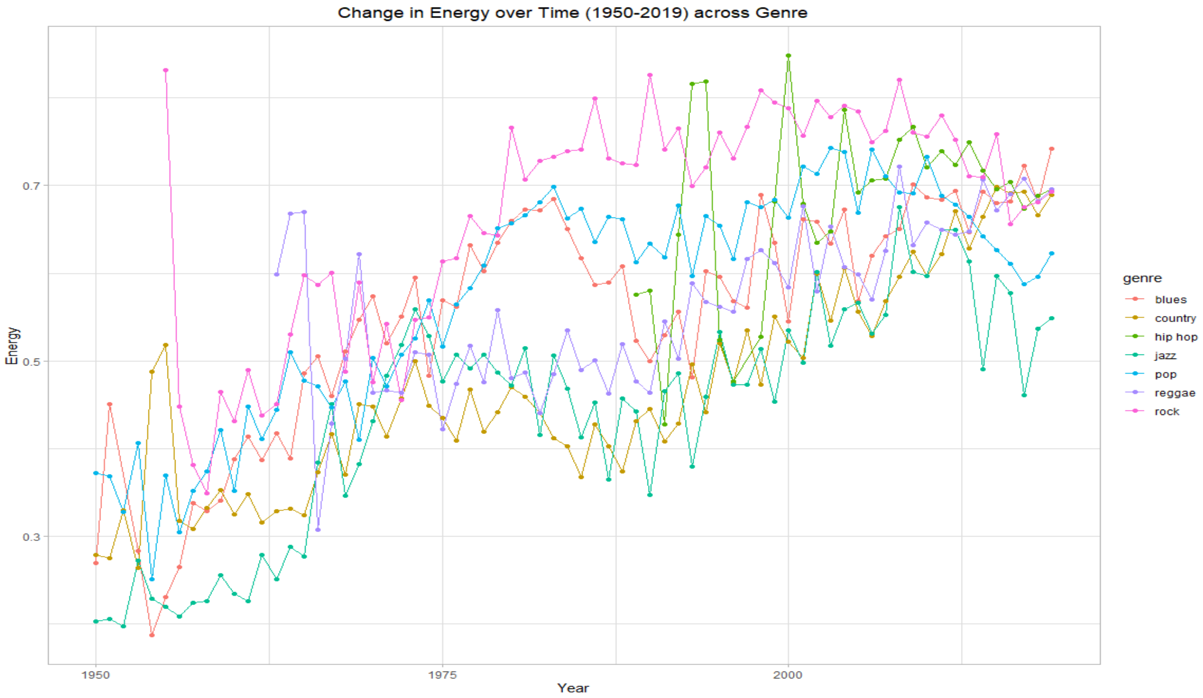
Across all genres except reggae, loudness has demonstrated an overall increase from 1950 to 2019. There was a drop in loudness of reggae music between the late 60’s and the early 90’s, but then loudness begins to climb again. The increase in loudness is most apparent beginning in the early 90’s which may be attributed to a shift towards louder music trends entering the 2000’s.



Acousticness has seen an overall decrease over time except for the hip-hop genre where there has been an increase. The decrease has not been linear though with many ups and downs over time. Jazz and blues saw a significant uptick in acousticness in the 90's and then went back down in the 2000's. The largest decrease in acousticness across genres can be seen in the early 1960's. The results make a lot of sense due to the prevalence of electronic instruments and synthesizers over natural acoustic sounds like the violin or guitar in recent times.



Energy has seen an overall increase in time for all genres except rock but again, it hasn't been a linear increase with many ups and downs over time. Rock saw a huge dip in the late 50's but has increased overall since then except a few more less significant dips in the 60's. This could be due to some sort of radical shift in trend or a musical revolution back in the day. Blues also saw a relatively huge hit in energy during the same period. The results do make sense overall as over time music has gotten more intense and active across the board.

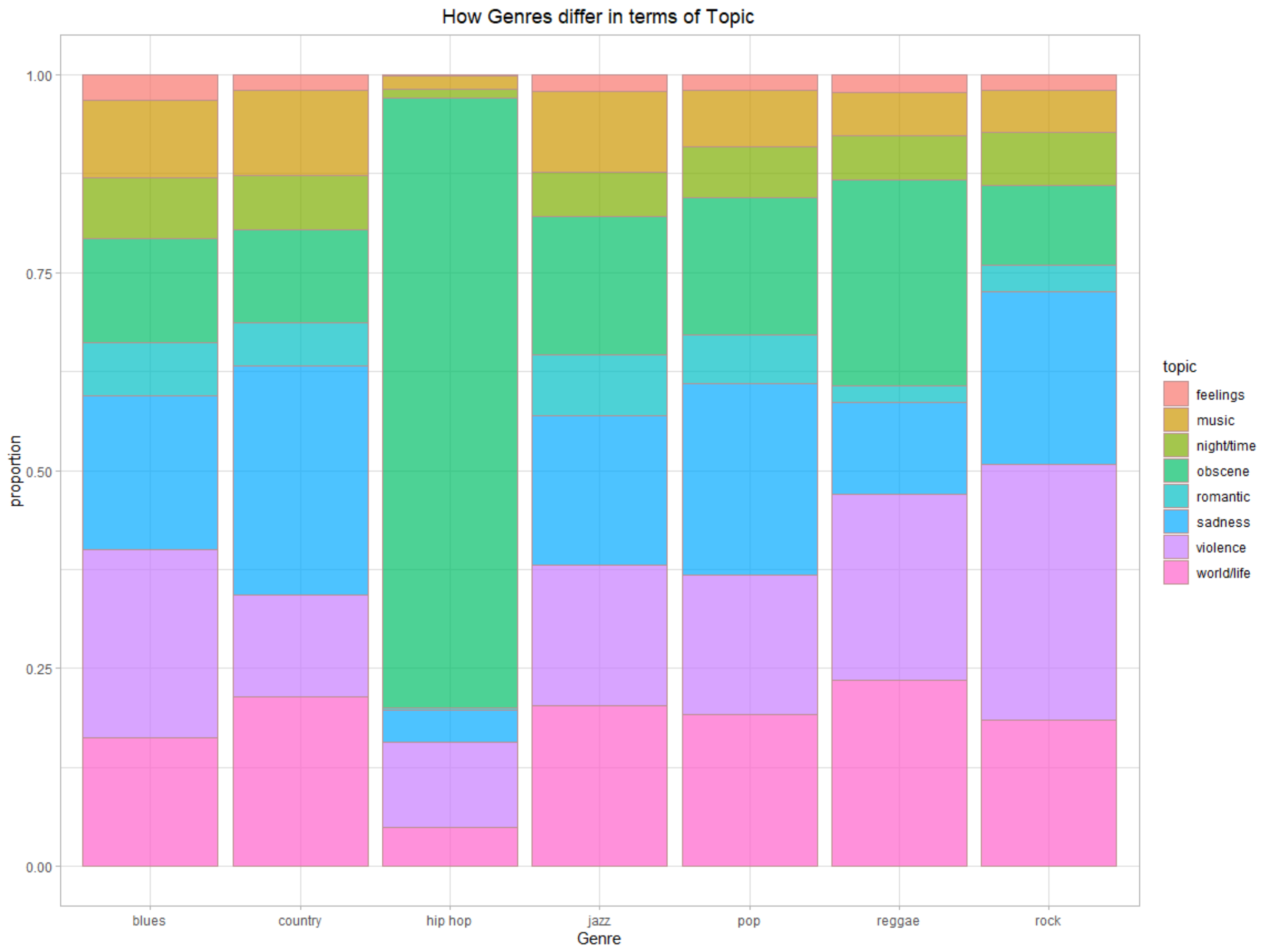


How The Genres Vary with Regards to The Songs' Content

We all listen to different songs and these songs belong to different genres of music. Some of the important discussions often brought forward whenever a new song is released centers around what type of message the song is trying to convey or whether the song is too lengthy or short. The next two sections will focus more on how the genres differ in terms of the message they are trying to pass along to potential listeners and how the songs are distributed by genre.

(1) How Genres Differ in Terms of Topic

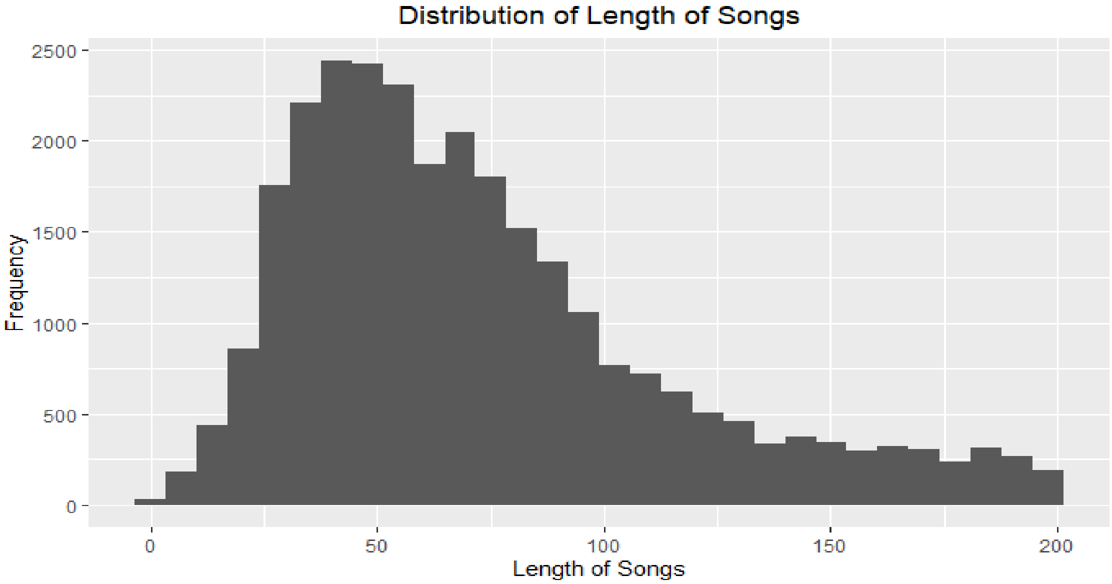
Another question of interest for this project was the type of message likely conveyed by the different genres. According to the chart below, blues songs are likely to promote violence. Country music is likely to be about sad topics. Hip hop songs are very likely to be obscene. This is not surprising as many hip hop songs tend to be vulgar. Jazz and reggae both have eclectic tastes. Jazz produces songs that either talk about life, promote violence or emit sadness. Reggae, on the other hand, focuses on life, violence and obscenity. Pop songs are mostly about sadness. Finally, rock songs promote violence. The different topics that are likely covered by songs in these genres indicate what topics most of their songs are likely to cover.

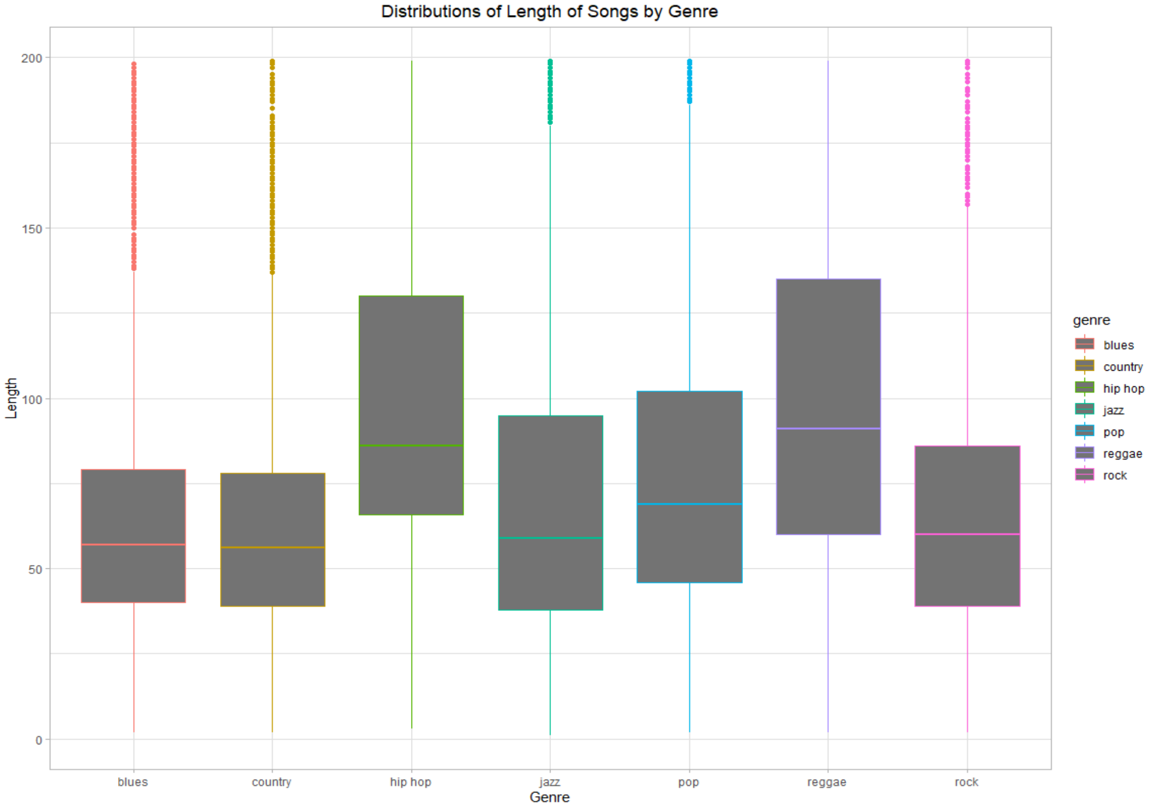


(2) Distribution of Songs' Duration & Distribution of Songs' Duration by Genre

The duration of music is also a major feature of music. In this data set, the length represents the occurrence of words instead of the length of the song in seconds or minutes. We can find that most lengths of music fall in around 40 to 50, and the length of the music is also spread out and reduced on both sides.

When divided by genre, we can find that hip hop and reggae have overall larger lengths than other genre types, and the next one is pop music. Blues, country, jazz and rock have relatively smaller length than the three discussed above.





Conclusions:

This project yielded numerous findings. According to the dataset, it was discovered that pop is the most popular genre amongst all the songs that were released by musicians from 1950 to 2019. A lot of musicians apparently release songs that fall under the pop category. Upon reviewing the relationship between valence and danceability across genres, there is no direct relationship between those two variables. However, it seems that people are more likely to dance to positive songs produced under the jazz genre.

Age was positively correlated with acousticness across all genres, and it was negatively correlated with energy and loudness across all genres. This implies that older songs were likely to be produced with more instruments and were designed to be lower energy. Energy and Loudness displayed the highest correlation between all the variables. Jazz accounted the most for this correlation as it had an energy-loudness correlation of approximately 0.8. Upon observing how the genres differ with respect to musical features, there were significant differences across genres. However, the sentiments emoted by the songs did not showcase any significant difference with respect to the genre. The phenomenon stated earlier about older songs lacking energy and displaying more usage of musical instruments can be witnessed in the line plots created during the research. Between 1950 and 2019, all the genres witnessed an increase in both energy and loudness, while they experienced a decrease in acousticness.

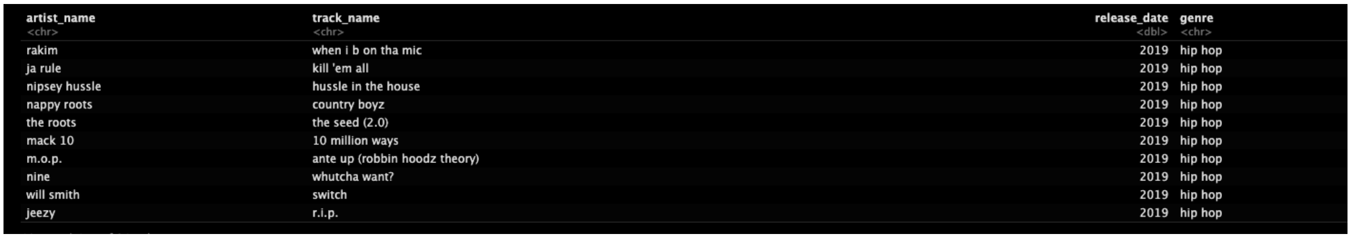
According to the topic covered by songs in each genre, jazz tends to produce songs about life, violence and sadness. Hip hop releases songs centered around obscenity. Blues and rock songs are likely to be violent. Country songs are fond of displaying sorrow. Reggae focuses more on topics like violence, life and obscenity. And finally, pop songs, like those of country, are likely to be sad as well as sharing messages about life.

The distribution of the duration of all the songs is positively skewed. However, when this distribution is investigated by genre. Blues, country, jazz, pop and rock remain positively skewed while reggae and hip hop are normally distributed.

Above all, this was a great project to learn about how musical genres differ across the ages with regards to parameters like energy, loudness and acousticness. It was also interesting to see how musicians in each genre are likely to produce songs covering certain topics while probably lasting for certain durations.

Limitations/Improvements:

While analyzing the musical data set, certain limitations made it impossible to properly understand certain relationships among variables. The release_date variable was quite inaccurate upon inspection as some old songs or songs created by dead musicians were claimed to have been released in 2019. This phenomenon can be observed in the image below:



artist_name <chr>	track_name <chr>	release_date <dbl>	genre <chr>
rakim	when i b on tha mic	2019	hip hop
ja rule	kill 'em all	2019	hip hop
nipsey hussle	hussle in the house	2019	hip hop
nappy roots	country boyz	2019	hip hop
the roots	the seed (2.0)	2019	hip hop
mack 10	10 million ways	2019	hip hop
m.o.p.	ante up (robbin hoodz theory)	2019	hip hop
nine	whutcha want?	2019	hip hop
will smith	switch	2019	hip hop
jeezy	r.i.p.	2019	hip hop

These songs are probably the remastered versions of the original songs. This happens quite often with a lot of songs as huge fans of these artists are driven to recreate the original versions of the songs. This makes it hard to have a justified idea about the age of the song. Another huge shortcoming was the data collector’s inability to properly document the meaning of the different variables, why he collected them and how they function. This prevents people using the data set from properly conducting rich forms of analyses. Some variables definitely had deeper meaning beyond their names. The solutions to the two issues cited above lies in the data collector properly signifying which songs are remastered as well as providing potential users of this data set with a verbose data dictionary describing the function of each variable.

The final issue with this project stemmed from not from the data set itself but rather the rules associated with it. R was the coding language that was strictly enforced by the professor to be used in this project. However, R’s sentiment analysis feature can analyze the sentiments emoted by words, not by sentences or paragraphs. Hence, this made it impossible to analyze the whole lyrics of the songs in the data set. Python, on the other hand, has a very robust sentiment analysis system that makes it possible for users to conduct sentiment analysis on sentences, paragraphs and lyrics. The instructor should have been a little bit lenient with this ruling so the lyrics in the data could have been optimally analyzed.