Studies of Audio Features

Elijah Muzzi
Washington & Jefferson College

Introduction:

The data being explored is Spotify audio features data collected from songs between 1960 to 2020. Exploration of this data may give us insight into what makes a “hit” song and how that “hit” recipe has changed over time. This type of data could be useful to artists when creating songs or, labels looking for an artist to pick up who they know can succeed. The data being analysed was scrapped from the Spotify API using a script. The Spotify API keeps information on different audio features of the song such as dancability, loudness, key, or speachinees, which a Spotify algorithm numerically defines. Some students from the University of San Fransico (Middlebrook, Kai, 2019) tried to create a model to predict hits by trying out several different types of models. They determined that Random Forest (RF) and Support Vector Machine (SVM) were the most accurate models and I hope to recreate or find a more accurate model in my exploration.

- Q1. How do audio features change over time?
  - Descriptive statistics

- Q2. Can we find different genres by clustering?
  - Clustering

- Q3. Can we see what audio features of songs when put together imply a hit?
  - Association Rules

- Q4. Which audio features of songs have the highest impact on being a hit?
  - Logistic Regression
  - Naive Bayes Classifier

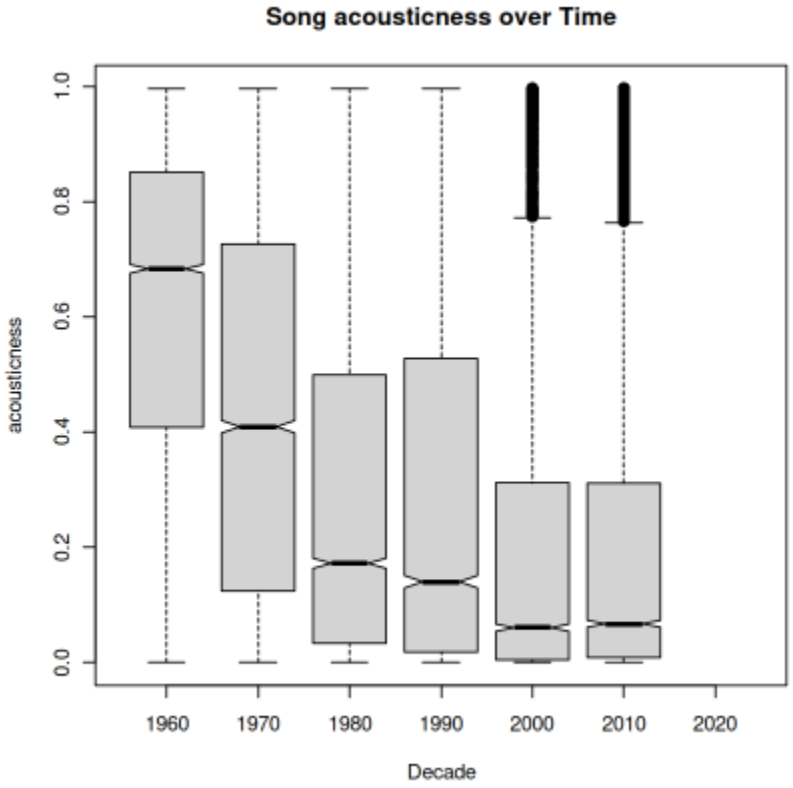
Data:

- Acousticness measures whether the track is acoustic.
- 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.
- 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".
- Key is what key the track is in. Integers map to pitches using standard Pitch Class notation. E.g. 0 = C, 1 = C♯/D♭, 2 = D, and so on.
- Liveness detects the presence of an audience in the recording.
- Loudness is the overall loudness of a track in decibels (dB). Loudness is the quality of a sound that is the primary psychological correlate of physical strength (amplitude).
- The Mode indicates the modality (major or minor) of a track, the type of scale from which its melodic content is derived.
- Speechiness detects the presence of spoken words in a track.
- Tempo is 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.
- Time signature (meter) is a notational convention to specify how many beats are in each bar (or measure). The time signature ranges from 3 to 7 indicating time signatures of "3/4", to "7/4".
- Valence is a measure 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).

Analysis:

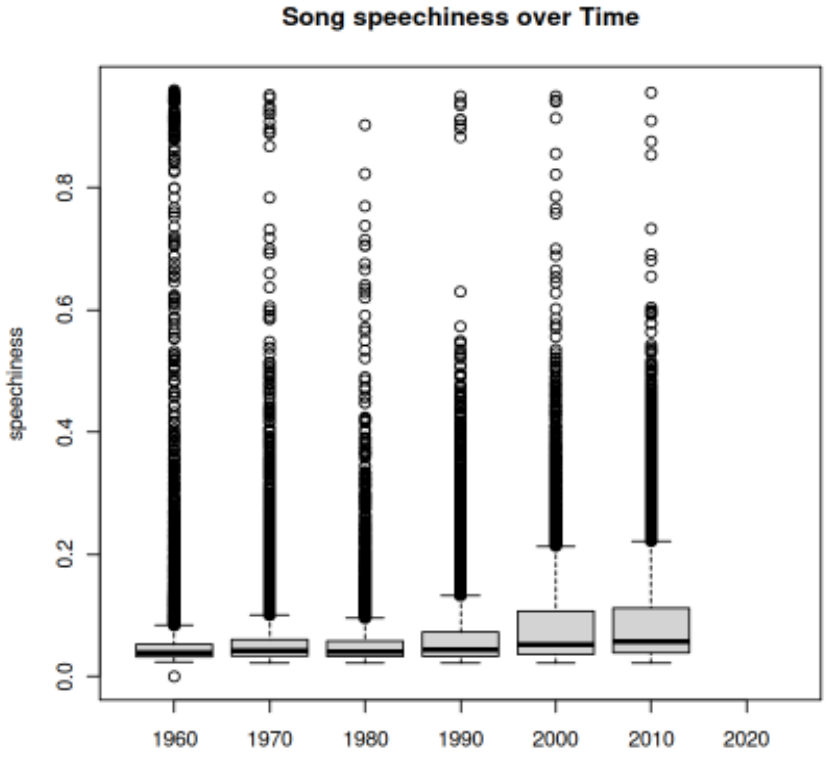
- **Q1. How do audio features change over time?**
    - Descriptive Statistics

Each row in the dataset represents the data collected for a specific song. There are 41,106 songs included in this dataset, spanning from 1960 to 2020. Below are some exploratory box plots looking at how the aspects of music change over time (Figures 1-10).



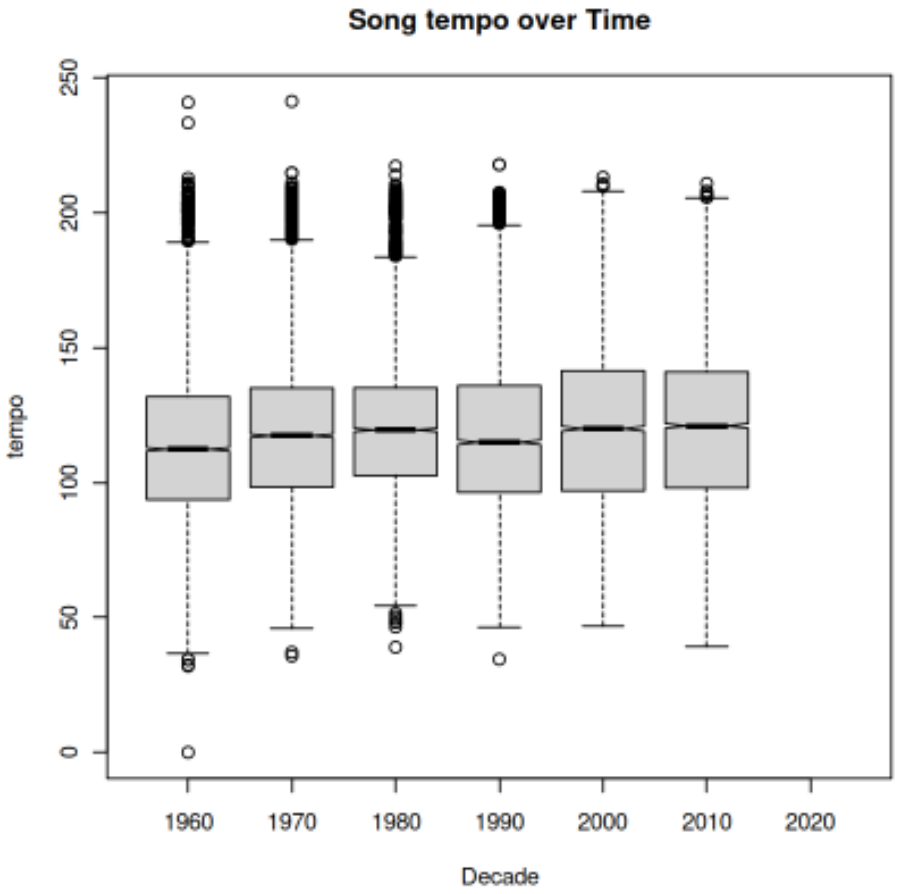
Over time we see a statistically significant difference in the mean values of acousticness as move forward in time. This may be a reflection of how electric instruments and production techniques have evolved. The outliers in the 2000s may represent a revival of interest in acoustic music.

Figure 1



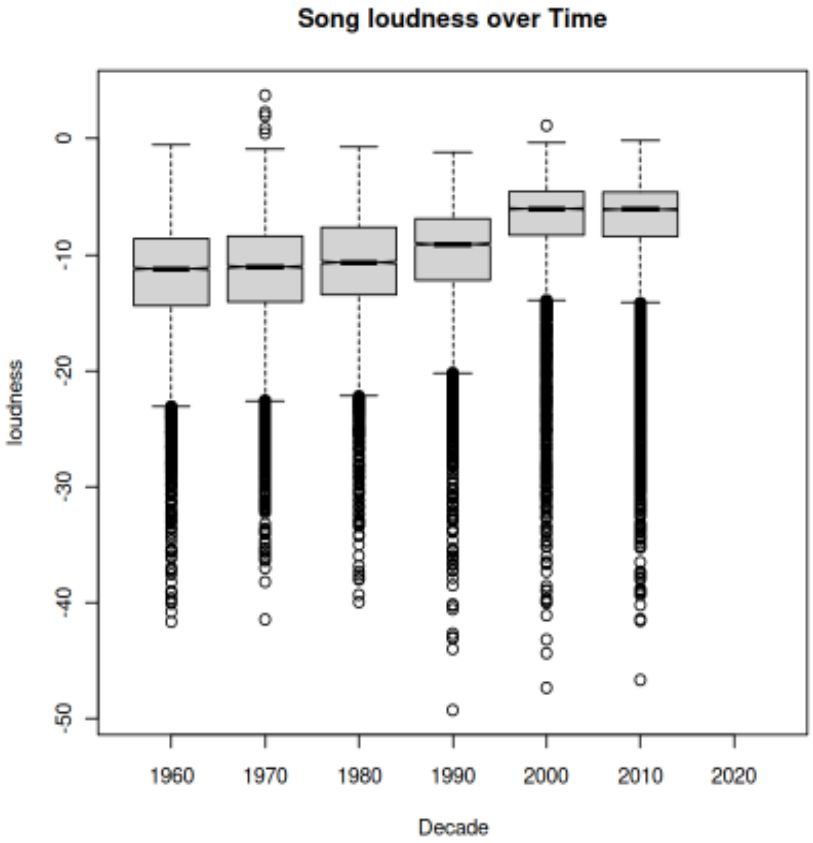
Over time we see an increase in the mean values of speachiness, which could be a reflection of a rise in popularity of hip-hop or other genres that are heavily lyrics focused. There are many outliers in every decade.

Figure 2



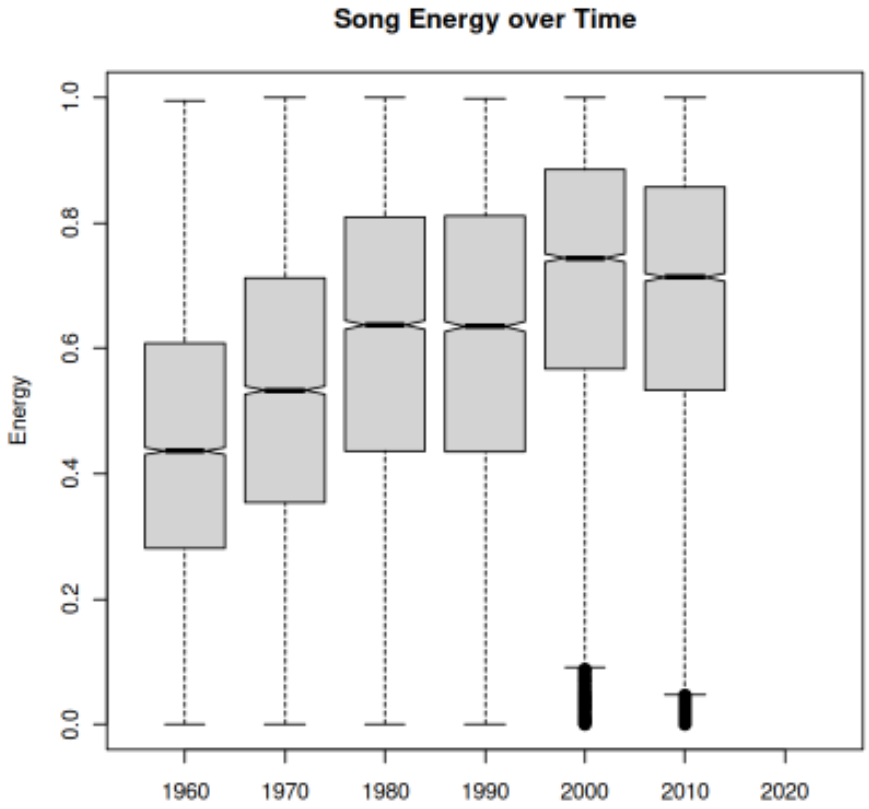
The mean song tempo has not changed much over time, the only decade that has a statistically significant lower mean tempo is the 1990s. Further exploration into this may yield interesting results.

Figure 3



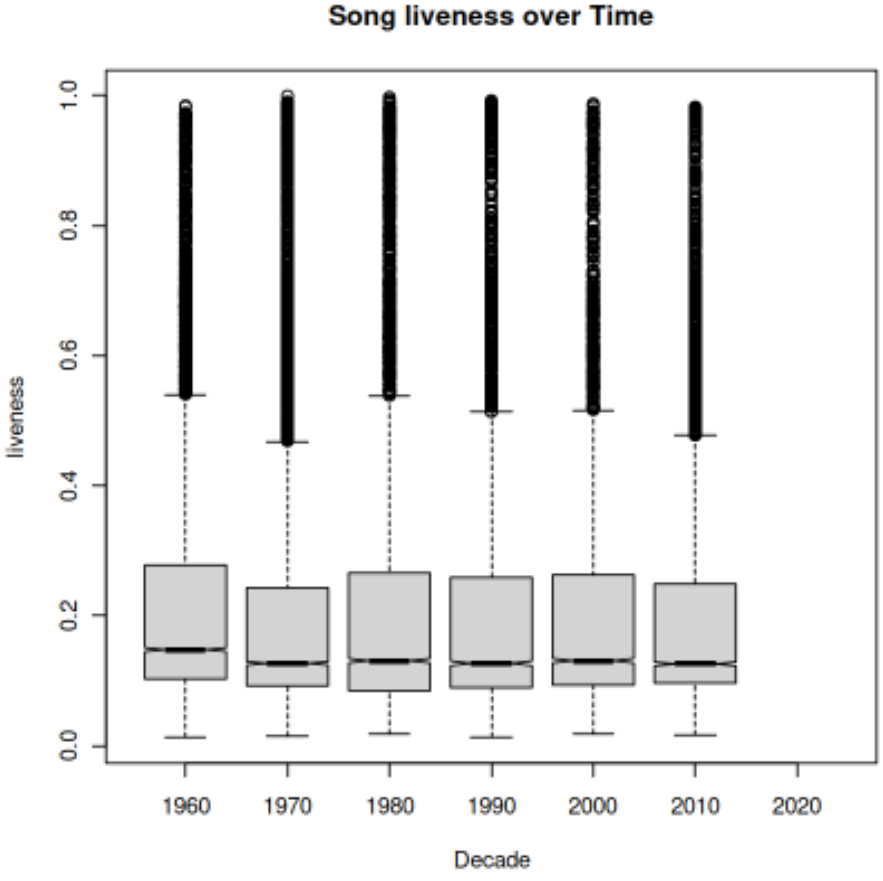
Loundess in songs has a statistically significant increase over the decades. The outliers in the 1970s may reflect beginning of the heavy metal era. The outliers in the 2000s may reflect the early dubstep era.

Figure 4



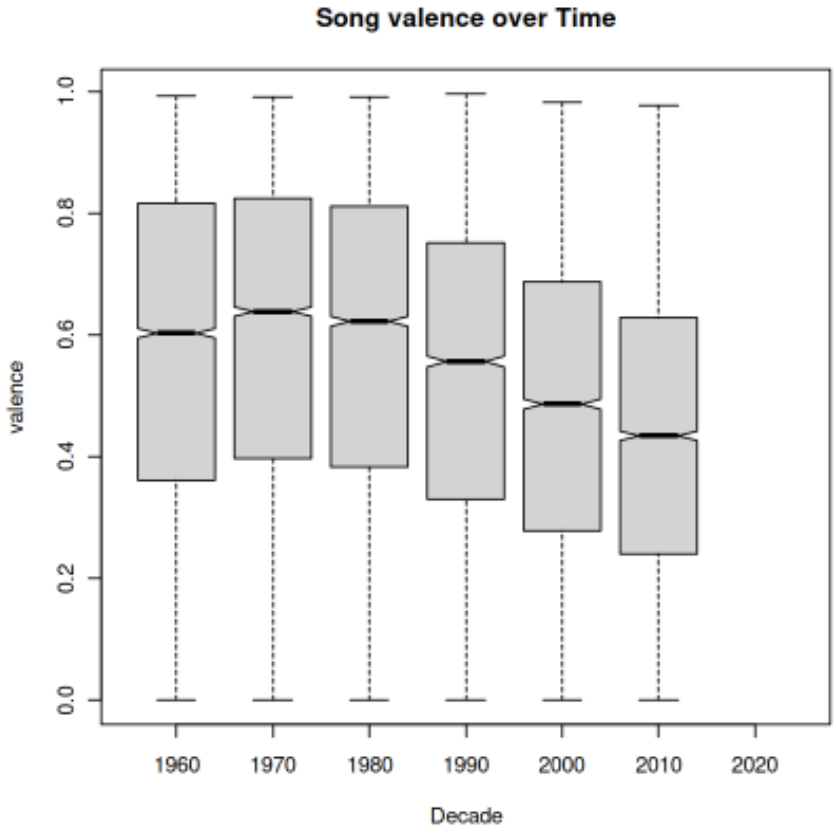
Song energy generally increases over time, with 1990 and 2010 being the two exceptions of a statistically significant decrease in mean song energy. It would be interesting to see how well mean song energy mirrors general happiness of the US population over the decades.

Figure 5



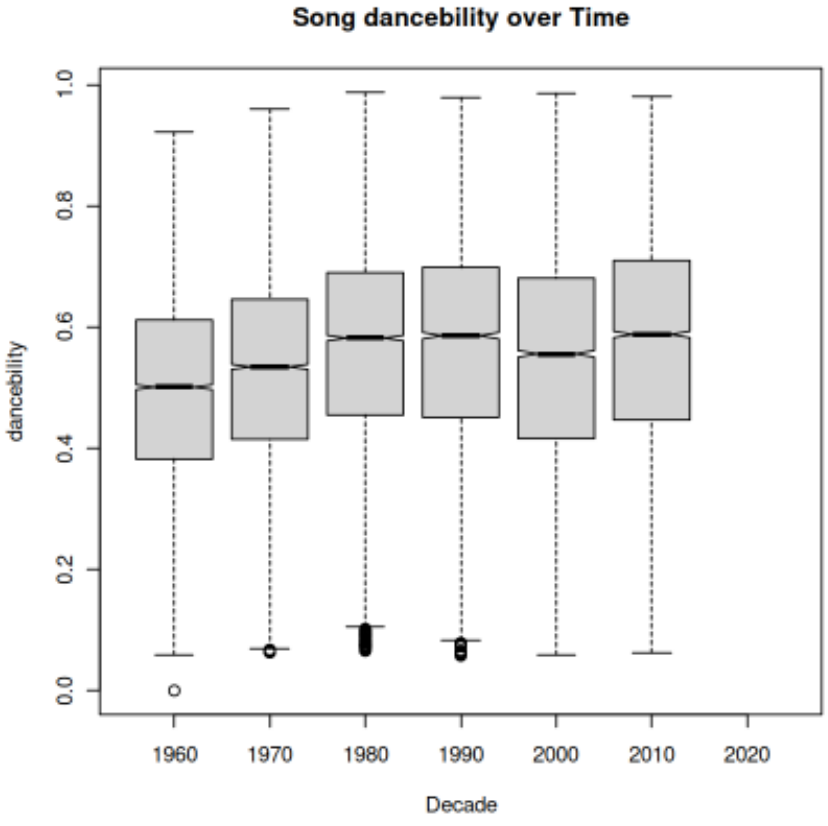
Mean song liveness seems to remain the same over the decades. Live performances will likely always been a large part of the music industry.

Figure 6



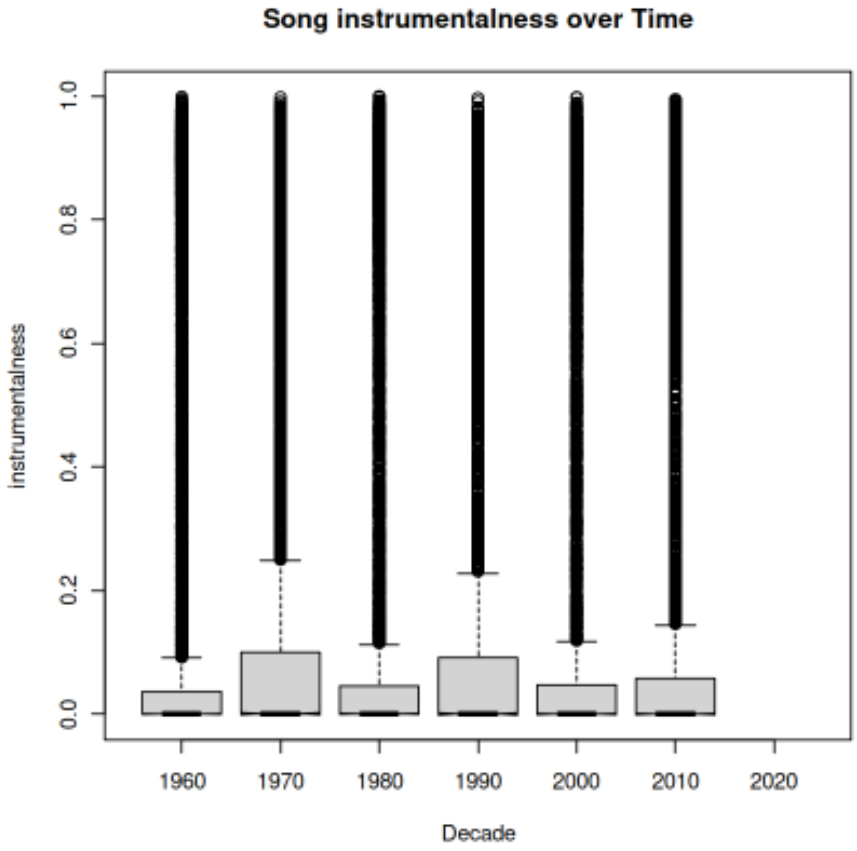
Mean song valence starts to have a statistically significant decrease as we enter the 2000s, implying a more negative mood overall in songs being released in those decades.

Figure 7



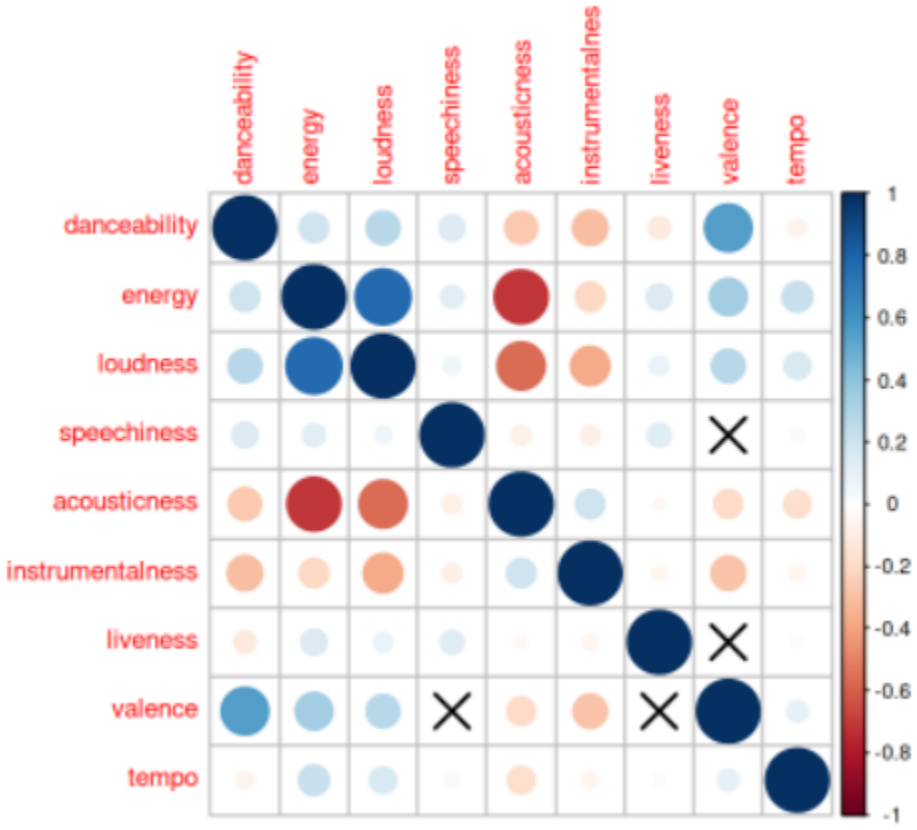
We see an upwards trend in song danceabilty over time, with the exception of 2000s.

Figure 8



Mean instrumentalness stays relatively constant with many outliers in every decade.

Figure 9



This is a correlation plot between all the numeric variables. Variables with a p-value greater than 0.05 are marked with an X. Valence is significantly correlated with speechiness and liveness

Figure 10

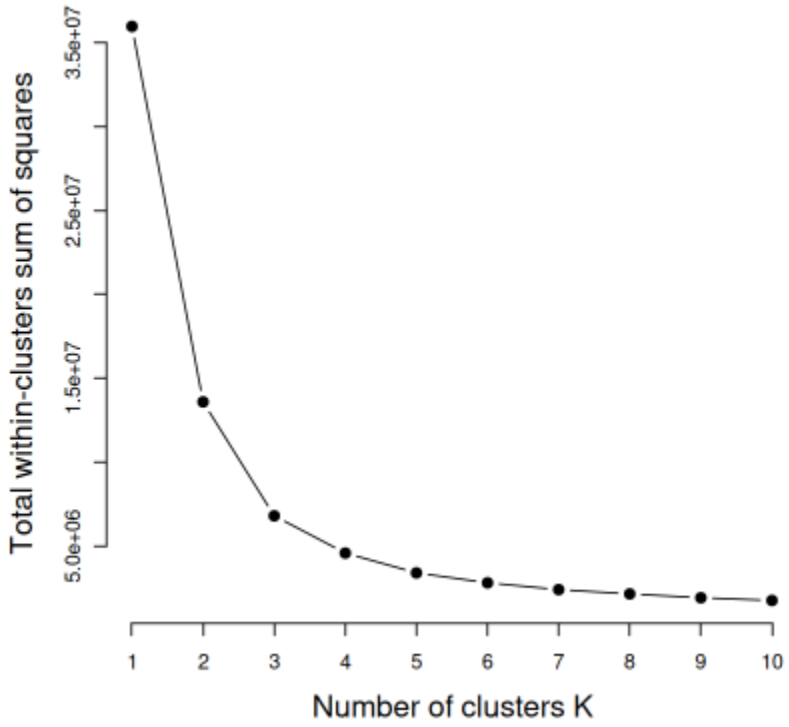
- **Q2. Can we find different genres by clustering?**
  - Clustering

The dataset, unfortunately, did not come with genres, so instead, the analysis will be utilizing time signature to predict genre. According to *Drums on Demand* genres of songs loosely follow this pattern

Time Signature	**Genre**
1/4	Other
3/4	R&B, country, ballads, scherzi, minuets, and waltzes.
4/4	pop, rock, funk, blues, country, and other Western genres.
5/4	Rock (especially progressive rock), Jazz, Classical, and other complex orchestral music use the 5/4 time signature.

Table 1. Time signatures and related genres.

The standard deviation between the different fields is great so the data was scaled before continuing.



Looking at this elbow plot, four centers seem to be where the dramatic decrease stops. Moving forward we will look at clusters based on four random centers. Using K-means clustering with 4 centers to look at how well the clusters it picked fit into the time signatures.

	Free Time	**1/4**	**2/4**	**3/4**	**4/4**	**5/4**
1	0	23	0	146	2129	87
2	0	72	0	603	17347	61
3	2	213	0	2218	7272	350
4	1	63	0	873	9549	97

Table 2. Songs in clusters 1-4.

Based on this table it would be appropriate to predict cluster 3 is part of the R&B, country, ballads, scherzi, minuets, and waltzes or, other genres. Cluster 2 is likely pop, rock, funk, blues, country, and other Western genres. It is hard to conclude which genres the remaining clusters would be composed of, this may imply Spotify aspects of songs may not be the best measure of genre or time signature.

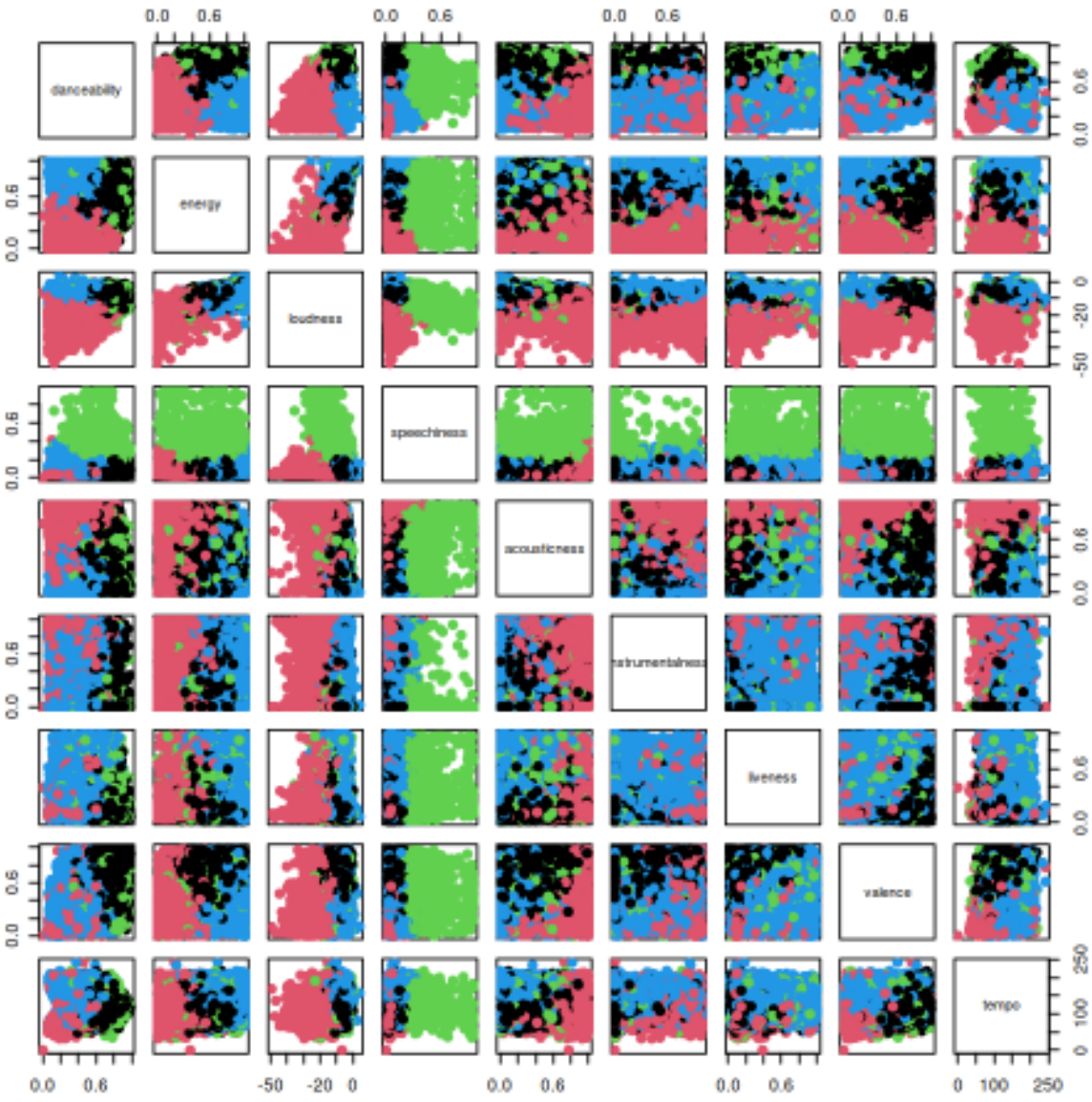


Figure 11. Plots of clusters found when comparing audio features.

Based on this pairs plot of the clusters we can see there is a lot of overlap between the 4 groups. Speechiness seems to be the variable with the most distinct four centers. Speechiness may be the best way to predict song genre.

- **Q3. Can we see what audio features of songs when put together imply a hit?**
  - Association Rules

To investigate which aspects of a song together imply a hit, a data frame was cleaned of everything but the numerical aspects of a song and turned each field into a boolean value whether or not it was high or low for that variable. A variable on the left-hand side of the rule implies a high value in that aspect. All ten rules have a lift over one so we know none of these rules occur due to random chance. The rule with the highest lift, rule 10, tells us that of the songs that were in the dataset, 23% of them had high danceability, energy, valence, and tempo. Out of those songs, 69% of them turned out to be a hit. Having high danceability, valence, and energy seems to be the best way to make sure your song is a hit.

```
     lhs                                      rhs      support   confidence
[1]  {danceability}                        => {target} 0.3762468 0.6188133
[2]  {energy, valence}                     => {target} 0.2721987 0.6389698
[3]  {danceability, valence}               => {target} 0.2841678 0.6394591
[4]  {danceability, energy}                => {target} 0.2921715 0.6809934
[5]  {danceability, tempo}                 => {target} 0.3762468 0.6188133
[6]  {danceability, energy, valence}       => {target} 0.2344913 0.6926062
[7]  {energy, valence, tempo}              => {target} 0.2721987 0.6389698
[8]  {danceability, valence, tempo}        => {target} 0.2841678 0.6394591
[9]  {danceability, energy, tempo}         => {target} 0.2921715 0.6809934
[10] {danceability, energy, valence, tempo} => {target} 0.2344913 0.6926062
     coverage  lift     count
[1]  0.6080134 1.237627 15466
[2]  0.4259962 1.277940 11189
[3]  0.4443877 1.278918 11681
[4]  0.4290371 1.361987 12010
[5]  0.6080134 1.237627 15466
[6]  0.3385637 1.385212  9639
[7]  0.4259962 1.277940 11189
[8]  0.4443877 1.278918 11681
[9]  0.4290371 1.361987 12010
[10] 0.3385637 1.385212  9639
```

Figure 12. Set of rules generated for a “hit” on the right-hand side.

Which audio features of songs have the highest impact on being a hit?

- Logistic Regression

To test what aspects of songs have the greatest impact on a song being a hit or not, a logistic regression model was created. The model took into account how danceabilty, energy, loudness, speechiness, acousticness, intrumentalness, liveness, valence, and tempo affect whether or not a song is a hit. The model found liveness and tempo to be the least statically significant having a p-value of 0.04. Below is the confusion matrix the model produced.

	Flop	**Hit**
Predicted Flop	2543	738
Predicted Hit	1487	3454

Table 3. Confusion matrix of predicted vs actual hits.

Based on this confusion matrix we know our model is correct 73% of the time which is good for this dataset because it is composed of 51% hits. Our model is more effective than predicting that every song is a hit or flipping a coin. Out of all the songs that turned out to be a hit our model successfully classified 82% of them, with only 18% of hits being a false negative. Out of all of the predicted hits, 70% of them turned out to be hits. If a music label invested in every predicted to be a hit in our model they would only lose money on their investment 30% of the time. The lowest accuracy measure in our model is the specificity, if a label has a gut feeling about a song that our model predicted to be a flop they would be right to trust their gut feeling 37% of the time.

- **Which audio features of songs have the highest impact on being a hit?**
  - Naive Bayes Classifier

The Bayes theorem describes the probability that A will occur if B has already been observed. Using the Bayes theorem on the same dataset was used in the logistic regression model to make predictions yielded the following confusion matrix.

	Flop	**Hit**
Predicted Flop	2237	569
Predicted Hit	1793	3623

Table 4. Confusion matrix of predicted vs actual (NBC).

The accuracy of this model is 71%. The Naive Bayes classifier is almost as accurate as the logistic regression model. The Naive Bayes is better at predicting hits and the logistic regression model was better at predicting flops. The Naive Bayes model might be better for an artist to use when they want to be certain which song is a hit and, the logistic regression model would be better for record labels to make safer investment decisions.

Conclusion:

Based on the analysis of this data set, I have found that while aspects of songs that Spotify's API collects are useful for predicting whether or not a song will be a hit through association rules and logistic regression. However, they are not that great at classifying songs under a specific genre or time signature. I could not find any other work using association rules to predict a hit however, Spotify uses association rules to feed its users song recommendations based on their taste. I was able to find two studies that tried to predict songs based on a Logistic model. The USF model had an accuracy of 81% and a precision of 74%, the Stanford model (Georgieva, Elena, 2018) had an accuracy of 74% and a precision of 72%, and my model had an accuracy of 73% and a precision of 70%. My model was not quite as good as the USF model but, about as good as the Stanford model. In my exploration, I found that speechiness is the best way to differentiate genre, high danceability, energy, valence, and tempo are part of 63% of hit songs and using Spotify data from songs I can use a Logistic Regression model to correctly classify 82% of hit songs, this model suits artist best because it allows them to catch the most hits. I would recommend the Naive Bayes classifier to record labels or managers looking for artists to pick up because you are less likely to get a false positive and can be more sure about your investment. In the future, I would like to use both neural networks and decision trees to see what other trends might exist in predicting hit songs.

Works Cited:

Middlebrook, Kai, and Kian Sheik. "SONG HIT PREDICTION: PREDICTING BILLBOARD HITS USING SPOTIFY DATA." Https://Arxiv.org/, 20 Sept. 2019, https://arxiv.org/pdf/1908.08609.pdf.

"Web API Reference: Spotify for Developers." Home, Spotify, https://developer.spotify.com/documentation/web-api/reference/#/operations/get-audio-features.

"How to Choose a Time Signature for Your next Song, Using Drum Loops." *Drums On Demand*, https://www.drumsondemand.com/blogs/news/15107149-how-to-choose-a-time-signature-for-your-next-song-using-drum-loops.

Georgieva, Elena, et al. "Hitpredict: Predicting Hit Songs ... - cs229.Stanford.edu." *STANFORD COMPUTER SCIENCE 229: MACHINE LEARNINGSTANFORD COMPUTER SCIENCE 229: MACHINE LEARNING*, https://cs229.stanford.edu/proj2018/report/16.pdf