ISYE 6740 Project Final Report

Group 36: Carmen Yu

December 2024

1 Problem Statement

If you were around in the 2010s and had stepped foot in a retail store or went to any party, there is a strong chance you heard one of Katy Perry's many popular singles from her album *Teenage Dream* blast through the speakers. Katy Perry's chart dominance in the 2010s was unparalleled and made her the first woman to have five No.1 singles from a singular album on the Billboard 100 charts.[6]

Yet in 2024, despite her previous success, Perry has failed to re-captivate audiences. Not only has her albums post-*Teenage Dream* failed to achieve equivalent success, her newest album *143* has been broadly panned. Most notably, critics have called her album 'dated' and that she has failed to evolve from her "2010s sound".[15][16]

Given those critiques, this brings up a broader question: **how much have themes and trends in music actually changed within the past decade?**

For this project, the objective is to use the Spotify audio features of songs from the year-end Billboard 100 chart from 2010 to 2023 to see how the trends from the previous decade have evolved to the current decade. Notably, trying to predict a 'hit song' with machine learning has been explored for quite some time now. It has turned into its own topic of interest called 'Hit Song Science'.[26] However, for this project, the interest is not solely on whether how the songs became a hit hit, but instead to see if machine learning can extrapolate what the current music themes and trends of popular songs to see if there are any musical commonalities between hit songs or if other external factors such as the level of fame of the artist would speak more to a song's success. In additional, the lyrics for the songs for each year will be examined to extract to see what themes and topics are popular.

1.1 Changes from Proposal

The project scope was adjusted from the original objective where there was an additional analysis on whether an artist's success dependent on evolving their sound to keep up with current musical themes and sounds by using Katy Perry's discography, specifically *143*, as a specific case study. This was of interest due to the critiques *143* received on being "dated" and how Taylor Swift has commented in her own documentary on how frequently reinvention is required for artists to continue to be relevant[28], something she herself clearly displays with her many album 'Eras'.

The adjustment was due to the fact that between the proposal and the final report, Spotify had decided to roll back the API access to the audio features data.[12] This made accessing any recent audio feature data impossible and limited the scope of the data available. Ultimately, this caused a pivot to use previous Spotify data from Kaggle instead. Since *143* was a very recently released album, this meant it was impossible to access and use *143* and the rest of Katy Perry's discography specifically for a case study.

2 Data Source and Preprocessing

2.1 Data Source

Data was assembled by using the *billboard.py* Python API to get the top hits of each year from 2010 to 2023.[5] Since the Spotify API limited the access to the audio features, the data for the audio features and lyrics were then assembled using a combination of 3 different Kaggle datasets.[3][4][1] Any missing lyrics was filled in using the Genius API with the help of the *lyricgenius* wrapper.[20]

Audio features are broken down into these categories in the following table. Descriptions are pulled from the Spotify documentation.[33]

Feature	**Description**
Acousticness	Confidence measure from 0.0 to 1.0 about whether the track is acoustic.
Danceability	Value between 0.0 to 1.0 for how danceable the track is. Based on tempo, rhythm stability, beat strength, and overall regularity.
Duration	Duration of the song, in ms
Energy	Value between 0.0 to 1.0 to measure of intensity and activity.
Instrumentalness	Likelihood from 0.0 to 1.0 to measure how likely the track contains vocal content.
Key	Integer value representing various pitches using standard Pitch Class notation.
Liveness	Likelihood value from 0.0 to 1.0, predicts the presence of a live audience.
Loudness	The overall loudness of the track in decibels (dB), averaged across the entire track.
Mode	Integer of either 0 or 1, representing the modality of the track. 0 = minor, 1 = major.
Speechiness	Value between 0.0 and 1.0 representing level of spoken word, where 1.0 represents tracks like audiobooks etc.
Tempo	The overall estimated tempo of the track in beats per minute (BPM).
Time Signature	Estimated time signature of the track.
Valence	Value from 0.0 to 1.0 that measures level of positivity in a track. Higher values are more positive.

Table 1: Spotify Audio Features Descriptions

2.2 Preprocessing

For the audio features, most values had a value between 0 and 1. However, a few categories such as loudness and tempo were in different scales/magnitudes. Hence, values were scaled using StandardScaler in the scikitlearn package for Python for the vector autoregressive models later on.

In general, any songs with missing audio features data were discarded from the overall data set. Imputation was considered, but due to the diverse nature of the values and the fact that this was less intended as predictive model and more of an analysis of trends, imputation felt like there was a strong chance of skewing the dataset and linear interpolation was not viable across songs. This did lead to a slightly imbalanced dataset, where more recent songs were under represented in the dataset due to data availability but this felt like a better trade off to imputation given the limitations.

For lyrics, basic text cleaning like removal of special characters and line breaks were carried out. The text was also transformed to lower case as well. Since any songs missing lyrics were collected via the Genius API, no considerations for imputation was needed.

3 Methodology and Evaluation

3.1 Audio Features

For the audio features, the features are analyzed using time series decomposition and vector autogressive models(VAR). Additional tests like the Augmented Dickey-Fuller (ADF) test and Granger Causality tests were also used in evaluating these models.

The usage of time series decomposition allows the extraction of the trend, seasonal and residual components of the Spotify data to analyze the change over time. This gives a better picture of how audio features has evolved over the past decade and what the current trends are and may be in the future.[14] Additionally, using a vector autoregressive model allows for a multi-variate time series model consideration where all the features can influence each other to change the outcome.

3.1.1 Time Series Decomposition

The audio features were decomposed using "Seasonal and Trend Decomposition using Locally estimated scatterplot smoothing (LOESS)" (STL decomposition). STL decomposition works by similarly to classic time series decomposition, but utlizes a LOESS smoother to fit the trend curve to allow more flexibility towards outliers.[11] It fits a trend curve using the LOESS smoother and then detrends the data to obtain the seasonality and then finally the residual. Essentially it breaks down the data into this form:

$$Y = T + S + R$$

where Y is the data and T, S, R are the trend, seasonal, and residual components respectively. The individual components are then graphed to see what trends and patterns emerge.

3.1.2 Augmented Dickey-Fuller Test

Before sending the data through a vector autogressive model, one of the requirements is that the data is stationary. Hence, it is likely necessary to detrend the data through differencing and use the ADF Test to test whether the data is stationary or not.[31]

3.1.3 Vector Autogressive Models

Vector Autogressive Models (VAR) works by construction a time series equation per feature that is a linear combination of its past values and past values of other features.[13] This allows insight to see if the different audio features have influenced each other and created any trends or correlated changes. A lag factor is included to determine how many past values to include into the equation. For example, for two features and a single lag, it would construct the following two equations:

$$y_{(1,t)} = c_1 + \alpha_{(1,1)} y_{(1,t-1)} + \alpha_{1,2} y_{(2,t-1)}$$

$$y_{(2,t)} = c_2 + \alpha_{(2,1)} y_{(1,t-1)} + \alpha_{2,2} y_{(2,t-1)}$$

where c is a constant, and α represent the coefficient of the equations. The coefficients tell us the impact of the lag value has on the current value of the feature. For example, if $\alpha_{1,2}$ is positive, it means the past value of feature 2 by 1 time order has a positive impact on the current time value of feature 1. Lag values are usually tested and determined using the lowest value for criterion such as the Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC).

3.1.4 Granger Causality Tests

It should be noted that the VAR model does not immediately mean that one feature ”causes” the other feature, but simply implies that there may be influence between them. Hence, an additional set of Granger Causality tests are also carried out. Granger Causality tests work by seeing if one time series can predict another time series by fitting various VAR models and then performs a series of F-tests to compare the models to see how well it predicts.[27] It essentially works by estimating how well X ”causes” Y, but it should be noted that true causality is hard to define and this test has a number of detractions.[27]

3.2 Lyrics

3.2.1 TF-IDF

TF-IDF is a way to encode words in text like the bag of words model. The difference is that TF-IDF takes into account of the frequency a word is used in comparison to the total amount of words in the entire text/corpus. This accounts for the ”term frequency” part of the model. The ”inverse document frequency” is calculated by taking the logarithm of the number of ”documents” the word has appeared in over the total number of ”documents”. In the case of this project, each ”document” would be a song. This is to balance more commonly used words that appear across multiple songs/documents. For example, if a word is used frequently in a single song but not a lot in other songs, it would probably be an important thematic word to that song. The final ”score” of the word would be the value of TF*IDF, which would represent the value of the word in the final encoding.[17]

TF-IDF was chosen over using bag of words for the text encoding as TF-IDF considers the context of the word against the entire corpus and offers a weighting score of each word, helping to highlight keywords better. Since lyrics have a natural structure of repetition and usage of similar words for thematic emphasis, TF-IDF will be able to using that structure to its advantage through its weighted scores.

3.2.2 K-means Clustering

After the words are vectorized through TF-IDF, this often produces a large sparse matrix with a high number of features due to the size of the corpus. Hence, the dimensions are reduced using SVD. This entire process is known as Latent Semantic Analysis.[7]

The reduced matrix is then passed through K-Means, and the usage of elbow plots are used to attempt to find an optimal k-value. After the final clustering, the silhouette score[10] could be used to check how well defined the clusters would be. A silhouette score closer to 1 indicates well separated and defined clusters, while a negative score indicates that incorrect clustering has occurred. Finally, the main key words are pulled out for each cluster to identify what major themes each cluster contains.

4 Exploratory Data Analysis

The histogram below shows the distribution of songs in the data across each year from 2010 to 2023.

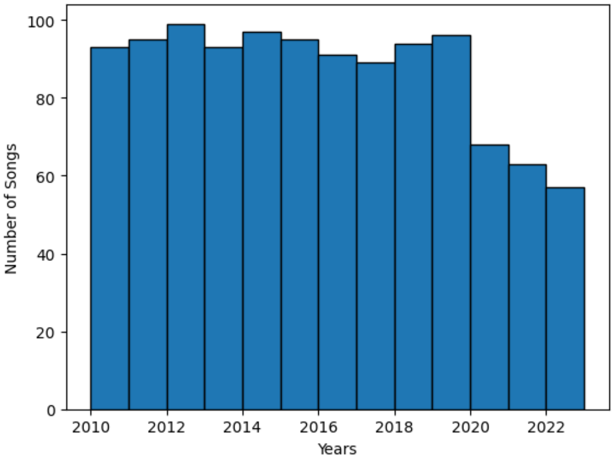


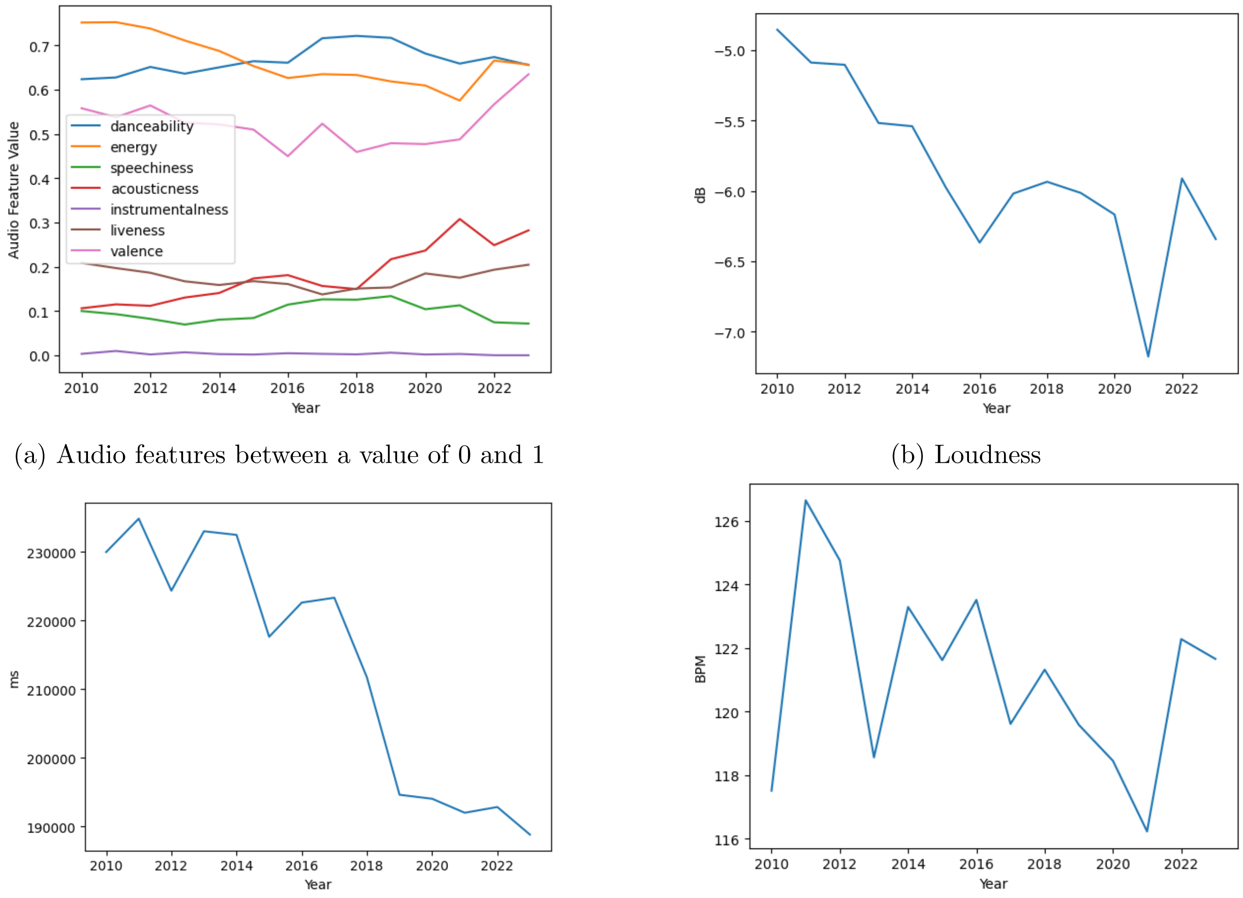
Figure 1: Histogram of songs per year

As it can be seen, songs for recent years are more under-represented, but this was expected due to preprocessing/data availability. All years from 2020 does have at least over half the amount of songs for any individual year in the 2010s. Ultimately, because the 2010s have a full decade of data, the data of the recent decade was always going to be somewhat under-represented.

4.1 Audio Features

For the audio features, it should be noted that key, mode, and time signature are categorical features that speak to a song's inherent musical construction. Since these features tends to be very specific and technical in nature, the exact impact of a song's key, mode, and time signature to an audience is rarely discussed. However, the other features such as danceability, energy, etc. are more mood descriptive and more likely to be used in general descriptions of the songs. Hence, for this project, these features were not utilized.

To look at the trend across the years, the data was grouped by each year and then averaged and plotted. Some features are on a different magnitude/scale, so those features were plotted separate on their original units of measure just to get an initial feeling on their change relative to a standard measure.



(a) Audio features between a value of 0 and 1

(b) Loudness

(c) Duration

(d) Tempo

Figure 2: Average value of audio features across each year

In general, for the audio features that have values between 0 and 1, we see a large spike in valence after 2020. This might speak to music moving towards a more positive direction post-pandemic. However, loudness and tempo also plummet after 2020 and pick back up after 2022, which suggests slower and softer songs was also in trend the same time valence picked up. This is interesting, as most people would likely associate more positive songs as more loud and upbeat. Most interestingly, duration had a large drop off after 2018, where the average song was around 20 seconds shorter. This phenomenon is likely due to the rise of Tiktok and music streaming.[18] It has been noted that streaming and Tiktok favour shorter songs, as shorter songs can have short snappy snippets that go viral and allow songs to rotate faster through Spotify to optimizing streaming payout for artists.

The features were then plotted onto a correlation plot to see if any notable correlations between features could be captured. Since there are already 10 features left, the idea would also be help to see what features would be helpful to examine to narrow down the scope and and focus of which features.

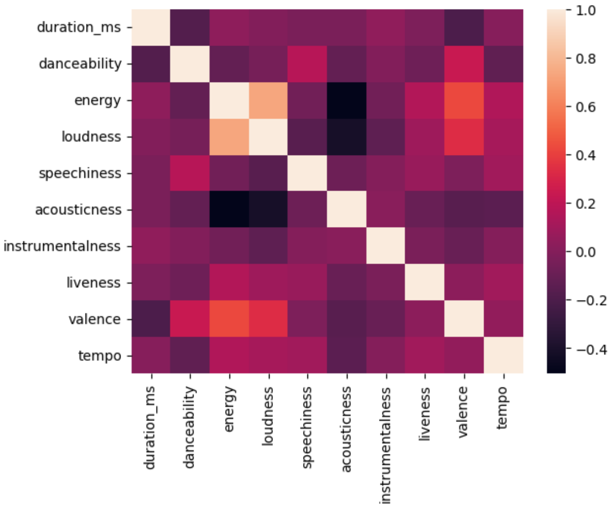


Figure 3: Correlation plot

The correlation plot above shows that some features that have stronger correlations are likely inherently linked such as energy, loudness, valence, and acousticness. This makes sense due to the definitions of those features, since they all speak to the how a song may be more upbeat or slower in nature. Interestingly, the correlation plot shows that there may be a negative correlation between the duration and danceability features, which may suggest that shorter songs tend to trend to be more ”danceable” for Tiktok.

Based on initial graphs and correlations, along with the definition of the features, the final features selected to be examined were acousticness, danceability, duration, energy, loudness, tempo, and valence. These features were selected for not only their individual trends, but their potential influence on each other's trends and to cover a broad range of what may make a song appealing to be a hit in the first place.

4.2 Lyrics

For lyrics, an initial look at some of the most common words for songs per year was examined. Stop words were removed, along with a subset of very frequently occurring words that had minimal meaning such as ”oh” and ”yeah”. An example for common words and its total count for the year 2010 is shown in the table below.

Word	**Count**
love	451
baby	322
want	273
say	261
low	240

Table 2: Most common words for 2010

Through out the analysis, the most difficult was to determine what additional words may be considered as stop words in the context of lyrics that may not be in the traditional set of stop words. The choice was made to err in favour of keeping more words vs eliminating more words in fear of losing context and because lyrics tend to be shorter and more repetitive than most documents analyzed in natural language processing.

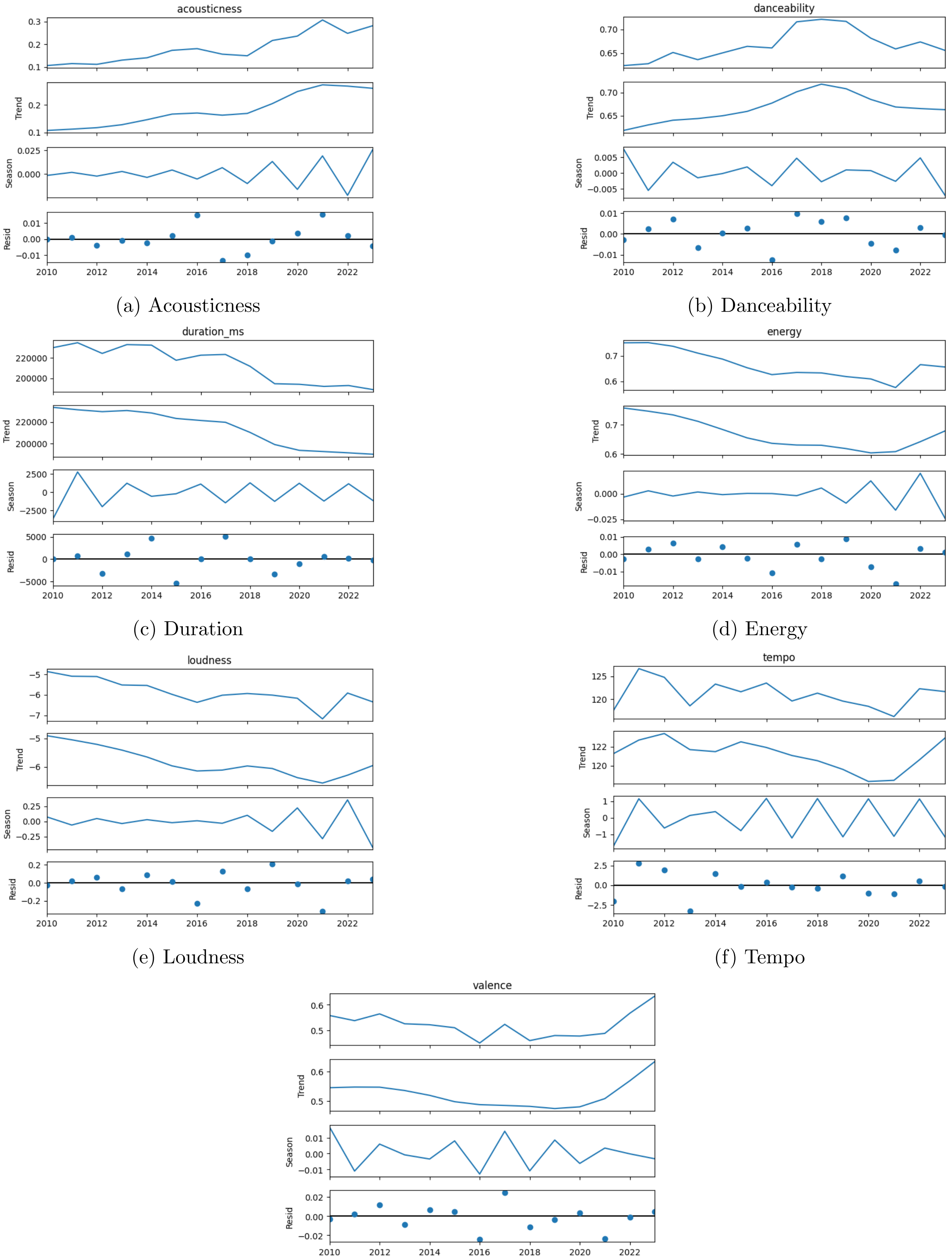
5 Results and Discussion

5.1 Audio Features

5.1.1 Time Series Decomposition

Since the data is split up into years, the values for the audio features were averaged to represent values for each year that could be decomposed. A period of 2 years was chosen, as some songs actually remained on the chart for about 2 years. Notably, only two songs were on the charts for repeated years. Both songs were Christmas songs, and one of them was the very famous Mariah Carey’s *All I Want for Christmas.* This indicates how strong the Christmas season influences the year-end chart.

The following time decomposition plots were generated for 7 audio features selected. The top graph shows the value, followed by the trend, seasonal, and residual graph. Note that since the features are independently decomposed, they have been left in their original units for easier interpretation.



(a) Acousticness

(b) Danceability

(c) Duration

(d) Energy

(e) Loudness

(f) Tempo

(g) Valence

Figure 4: Time Series Decomposition

From the graphs, we see a notable trend downward in duration and an upward trend in valence. Acousticness has also risen significantly over the years, especially starting around 2018. This indicates a rise in acoustic songs being popular with the general audience. At the same time, danceability appears to peak around 2018 and decrease slightly going into the new decade, which may be related to the acoustic factor rising around 2018.

Tempo also seems to be trending upward for this current decade, but interestingly, the seasonal factor of tempo indicates that there may be some level of seasonality of uptempo and downtempo songs across the years. None of the other factors have particularly strong seasonal trends, but since 2015, tempo shows a pattern of increasing and decreasing. It appears there might be a trend that when downtempo songs eventually are the majority of the popular songs, listens might be craving uptempo songs which then eventually find their way back onto the charts.

There also appears to be some potential seasonality in duration, but in general it seems to trend downward regardless.

Looking at the residual plots, they generally seem to be centered around 0, but there does seem to be some level of potential cyclic trend in the noise for some of the features, such as danceability and valence. This indicates there may be some either broader or narrower seasonality trend that has not been picked up by the decomposition. This could potentially be captured by experimenting with different period lengths with a broader set of songs at a more frequent time period than years. (e.g. the monthly charts instead of year-end.)

5.1.2 Augmented Dickey-Fuller Test

Since VAR models operate on an assumption that the data is stationary and it was clear there was definitely some level of trend within the audio features, the ADF test is performed to see which features potentially need differencing to detrend the data. The following p-values are obtained after the test for each feature and evaluated against a confidence level of 0.05.

Feature	**p-val**	**p <0.05**
danceability	0.434	False
valence	0.725	False
acousticness	0.996	False
loudness	0.042	True
energy	8.347e-05	True
tempo	0.004	True
duration	0.910	False

Table 3: ADF results

Notably, over half the features fail. This indicates the danceability, valence, acousticness, and duration has strong trends over the years. After a round of first-order differencing, the following ADF p-values are achieved.

Feature	**p-val**	**p <0.05**
danceability	0.030	True
valence	0.996	False
acousticness	0.002	True
loudness	0.042	True
energy	8.347e-05	True
tempo	0.004	True
duration	0.001	True

Table 4: ADF results after differencing

All values but valence are now sufficiently de-trended. Additional testing found that even further differencing, valence would still fail the test. Hence, the decision was made to leave valence out of the VAR models for this project. Other forms of either differencing or models could be considered for the future to potentially analyze the influence of valence on other features.

5.1.3 Vector Autogressive Models

Referring back to the correlation plot, two VAR models was selected. One model had acousticness, loudness, and energy while the other had danceability and duration. These pairings were chosen due to the correlation plot hinting at potential relationships in these features.

For the first model with acousticness, loudness, and energy, the lag order was determined to be 2. Along with AIC and BIC, the model also estimates with Final Prediction Error (FPE) and Hannan-Quinn Information Criterion (HQIC) as additional criterion metrics of which lag order is potentially the best.

The result for the first VAR model are shown below, where each feature has their corresponding coefficients with their lagged terms and corresponding p-values.

		Acousticness		**Loudness**		**Energy**	
	Terms	Coefficient	p-val	Coefficient	p-val	Coefficient	p-val
Lag	constant	0.310	0.079	0.101	0.603	-0.057	0.712
t-1	acousticness	-0.462	0.268	0.107	0.816	0.217	0.551
	loudness	0.851	0.035	-0.972	0.030	-1.311	0.000
	energy	-0.009	0.990	1.090	0.164	2.024	0.001
t-2	acousticness	0.222	0.459	-1.249	0.000	-0.449	0.086
	loudness	1.572	0.053	-0.841	0.350	-0.131	0.854
	energy	-1.828	0.057	0.779	0.464	-0.158	0.851

Table 5: VAR Model 1 results

As it can be seen, most of the coefficients' p-values are quite large. In this case, a p-val with less than 0.05 confidence level would say that the lagged term has an impact on the feature's present value.

For acousticness, it seems like a first order lagged term of loudness may have a positive impact. It suggests that perhaps if songs were loud in the past, it may increase the level of acousticness in songs in the future.

For loudness, it seems like a first ordered lagged term of loudness and a second ordered lagged term of acousticness has negative impact. Hence suggesting that if songs were loud just recently, it may trend softer, but also that if softer songs were just in trend, it may also decrease how loud the songs may be.

Finally, for energy, it seems like a first order lagged term of loudness and energy may have a negative and positive impact on its value. Interestingly, it seems to suggests that if songs were energetic in the past, it would stand to continue to rise, but would get softer if past songs were louder. Since people generally associate louder as more higher energy, this seems to suggest some separation of how the two values should be interpreted.

For the second model with danceability and duration, a lage order of 2 was also used. AIC, BIC, and FPE had a hard time determining an appropriate lag order as all three criterion suggested a lag order of 0, meaning that these features may not have been well versed for a VAR model. However, HQIC did end up suggested a lag order of 2. Hence, a lag order of 2 was chosen to see how the results panned out. The following table show the coefficients and their related p-values.

		Danceability		**Duration**	
	Terms	Coefficient	p-val	Coefficient	p-val
Lag	constant	0.314	0.345	-0.347	0.123
t-1	danceability	-0.051	0.909	0.046	0.879
	duration	1.141	0.032	-0.326	0.364
t-2	acousticness	0.265	0.432	-0.225	0.323
	loudness	0.318	0.661	-0.489	0.317

Table 6: VAR Model 2 results

Looking at the p-values, it can be immediately seen why this model had a hard time. The only coefficient that passes the threshold (and any reasonable threshold at all) is the first ordered lagged term of duration for danceability. It seems like duration would have some positive impact on danceability, where a longer song would be more danceable. Overall, while the correlation plot had indicated potential correlation, there is minimal influence of these features on each other.

5.1.4 Granger Causality Tests

To further check on the confidence of the VAR results, Granger Causality tests are performed on the various pairs. This will help see the level of confidence that these features can actually be used to predict the other feature. Note that a series of four tests were carried out here for each pair of features at each lag order, but all four tests return similar results. For conciseness, the F-test will be the main test of interest here. As well, the order of the features matter here as the test is a one way test, seeing if X ”causes” Y, but does not examine if Y ”causes” X.

The p-values obtained the test for each pair is shown in the table below.

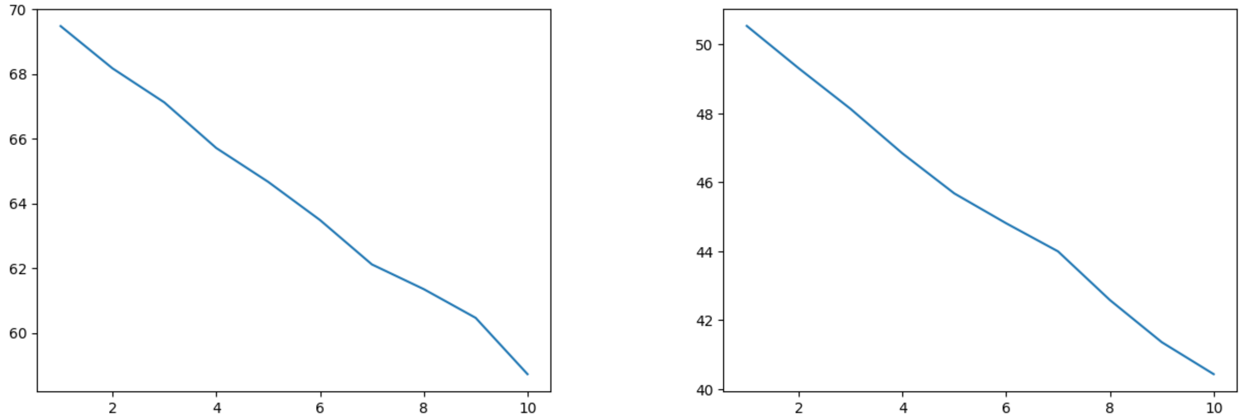
	p-val	
Lag	1	2
(loudness, energy)	0.0217	0.2170
(energy, loudness)	0.0009	0.0208
(loudness, acousticness)	0.5349	0.2209
(acousticness, loudness)	0.3707	0.3299
(acousticness, energy)	0.9727	0.9209
(energy, acousticness)	0.4462	0.8636
(danceability, duration)	0.0326	0.1810
(duration, danceability)	0.3887	0.6338

Table 7: Granger Causality F-test results

The results show that despite the VAR model results, the level that each feature can truly help predict the other features may be somewhat limited. Almost all the p-values are greater than 0.05, with the exception of energy and loudness along with the first ordered lagged term of loudness for energy and danceability for duration. Hence, this indicates that while some level of relationship may exist between these features, this does not mean they are particularly strong at predicting the values for each other. This is interesting as it might mean that while a hit song model could capitalize on any pre-existing relationships between the features that may exist, it may have a hard time in reality predicting for future hits.

5.2 Lyrics

For lyrics, after the initial look at the most frequent terms, the terms were passed through the TF-IDF vectorizer and reduced via LSA. Elbow plots were then generated to attempt to find an optimal k-value. Since clustering was carried out for every year, for conciseness, only 2 elbow plots are being shown below.



(a) 2010

(b) 2020

Figure 5: Elbow Plots for 2010 and 2020

Notably, there is very little elbow in these plots. This is actually the case for pretty much every year. Even when the range of the k-value was expanded, the plots just showed a downward trend in general. This is an initial sign that clustering may not be the best choice here. Ultimately, a choice of around 7 - 10 clusters was chosen for each year here. A table for the silhouette scores for each year is shown below.

Year	**Silhouette Score**
2010	0.012
2011	0.014
2012	0.021
2013	0.021
2014	0.014
2015	0.012
2016	0.035
2017	0.030
2018	0.054
2019	0.041
2020	0.015
2021	0.043
2022	0.124
2023	0.169

Table 8: Silhouette scores for each year

From the table, it can be seen that all of the silhouette scores, while positive, are extremely close to 0. This indicates that the clusters are all extremely close to each other and/or overlapping each other. This likely explains why the model had a hard time determine the level of clusters through an elbow plot as it indicates that the data is all relatively similar, and increasing the number of cluster just meant creating smaller clusters of close data points which would lead to a lower within-cluster squared distance.

This is noticeable in the keywords as well. For example, here is a table for the main keywords extracted for the year 2010 for each cluster.

Cluster	**Keywords**
1	tonight baby want love boy way come night magic bring
2	love break want heart low gone tell gotta somebody falling
3	daddy soul dream way believing home maybe wanna tonight thought
4	head going song ay right stuck smile ropes shawty chorus
5	life win time thinkin honey better long em sky laughing
6	say baby gon girl need nothin way stuck tell hard
7	chick bad wish sexy ha damn girl answer use right

Table 9: Keywords for 2010 clusters

The cluster themselves generally seem somewhat similar to each other and seem to contain some general words that doesn't inherently relate to each other. Across the years, the general themes that seem to occur in the clusters were: partying, love, and lust. References to money, driving, drinking, and dreaming were quite frequent, and clusters would occasionally catch smaller themes such as strong usages of NSFW words, Spanish if there was a strong presence of Latin music that year on the charts, and Christmas theme words if Christmas songs was doing well that year on the charts.

In general, it does not appear that themes have really at all shifted all that much across the years. The words in lyrics continue to remain similar and difficult to isolate independent themes through K-means regardless of the year. This might indicate how most of the time, song lyrics use a lot of the same words in different ways and hit songs generally tend to hit very universal and classic themes that are relatable to a large majority that are timeless in nature.

The results also might indicate that perhaps TF-IDF and K-means was just not the appropriate model for the job. Given the assumption above, it appears that the *structure* and the way lyrics are phrased in the song may matter more for extracting broader themes. The downside of TF-IDF is that it loses contextual meaning as it removes sentence structure from the data and this may conclude that song lyrics require a model that can pick up contextual meaning to truly analyze broader themes.

6 Conclusion

Overall, the project highlighted on how audio features and lyric of hit songs may have changed over the years and their potential trends and influences on each other. However, it is important to note that some of the findings seem to indicate that the audio features and lyrics do not seem very predictive of how hit songs may evolve in the future. That is, the way hit songs are right now do not seem to give any strong indications of how hit songs *may be.* For example, while it was clearly seen that a duration of a song had a downward trend, the *cause* of that trend is likely external to the song's composition, as Tiktok and steaming likely created that influence. The VAR models and Granger Causality tests did not show that any past values of other audio features had any major influence on duration. Hence, this raises this question: is it worth trying to build hit song models based on audio features and lyrics at all? And if so, what external factors should be used instead to measure hit song success?

In addition, this project was focused on using ”hit songs” from the year-end Billboard 100 chart. The question of whether the Billboard 100 is a good representation of a ”hit song” is also worth considering in this context. Billboard themseleves have said that fans of certain artists work together to push up the chart numbers of songs that the artist releases so they can have greater charting success.[19] At that point, is the song's success on the charts a reflection of the quality of the song making it popular or is it a reflection of the popularity of the artist to begin with? Al-Beitawai et.al (2018) also considered at the end of their study for their hit song model to add a factor for whether or not the song was of an established artist, as they suspected it may influence whether the song was a hit or not.[2]

6.1 Future Considerations

Notably, the data for this project was primarily focused on the English-speaking and North American market. Some potential expansions would to be include the influence of the market for the data.

It would also be interesting to carry out the exact same procedure on *non-hit songs*. If the results for non-hit songs are similar to how hit-songs behave, it would truly indicate that perhaps the audio features and lyrics aren't really the deciding factor for what makes a song pop on the charts.

As well, a redo with a broader range of data where the songs were captured on the monthly charts instead may also provide a more granular look at the trends and potential seasonality. This would be useful especially since holiday songs, such as Christmas songs, do seem to strongly influence the charts.

References

[1] *960K Spotify Songs With Lyrics data*. Aug. 2024. URL: https://www.kaggle.com/datasets/bwandowando/spotify-songs-with-attributes-and-lyrics/data.

[2] Zayd Al-Beitawi, Mohammad Salehan, and Sonya Zhang. "What makes a song trend? Cluster analysis of musical attributes for Spotify top trending songs". In: *Journal of Marketing Development and Competitiveness* 14.3 (2018). URL: https://articlearchives.co/index.php/JMDC/article/view/4610/4572.

[3] *Billboard Hot weekly charts*. Dec. 2023. URL: https://www.kaggle.com/datasets/thedevastator/billboard-hot-100-audio-features.

[4] *Billboard Hot-100[2000-2023] data with features*. June 2024. URL: https://www.kaggle.com/datasets/suparnabiswas/billboard-hot-1002000-2023-data-with-features.

[5] *Billboard.py*. Aug. 2024. URL: https://pypi.org/project/billboard.py/.

[6] Eric Renner Brown. "Katy Perry's 'Teenage Dream'; defined a pop era - but also marked the end of one". In: (Aug. 2024). URL: https://www.billboard.com/music/pop/katy-perry-teenage-dream-era-greatest-pop-stars-1235757714/.

[7] *Clustering text documents using k-means*. URL: https://scikit-learn.org/1.5/auto_examples/text/plot_document_clustering.html#k-means-clustering-on-text-features.

[8] Ioannis Dimolitsas, Spyridon Kantarelis, and Afroditi Fouka. "SpotHitPy: A Study For ML-Based Song Hit Prediction Using Spotify". In: *arXiv.og* (Jan. 2023). URL: https://arxiv.org/pdf/2301.07978.

[9] Juan Du. "Sentiment Analysis and lyrics theme recognition of music lyrics based on natural language processing - ProQuest". In: *Journal of Electrical Systems* 20.9 (2024). URL: https://www.proquest.com/openview/3481614d7f3825c8cb35d9b3a46f1c9a/1?pq-origsite=gscholar&cbl=4433095.

[10] GeeksforGeeks. *Clustering metrics in machine learning*. Mar. 2024. URL: https://www.geeksforgeeks.org/clustering-metrics/#silhouette-score.

[11] Rob J Hyndman and George Athanasopoulos. *6.6 STL decomposition*. 2nd ed. 2018. URL: https://otexts.com/fpp2/stl.html.

[12] *Introducing some changes to our Web API*. Nov. 2024. URL: https://developer.spotify.com/blog/2024-11-27-changes-to-the-web-api.

[13] Rob J Hyndman and George Athanasopoulos. *11.2 Vector autoregressions*. 2nd ed. Melbourne, au: OTexts, 2018. URL: https://otexts.com/fpp2/VAR.html.

[14] Rob J Hyndman and George Athanasopoulos. *6.1 Time series components*. 2nd ed. Melbourne, au: OTexts, 2018. URL: https://otexts.com/fpp2/components.html.

[15] Maura Johnston. "Katy Perry's '143'; Is a Failed Attempt to Rekindle Her Glory Years". In: (Sept. 2024). URL: https://www.rollingstone.com/music/music-album-reviews/katy-perry-143-review-1235108317/.

[16] Rich Juzwiak. *Katy Perry: 143*. Sept. 2024. URL: https://pitchfork.com/reviews/albums/katy-perry-143/.

[17] Fatih Karabiber. *TF-IDF — Term Frequency-Inverse Document Frequency – LearnDataSCI*. URL: https://www.learndatasci.com/glossary/tf-idf-term-frequency-inverse-document-frequency/.

[18] Elias Leight. "Here's why shorter songs are surging (And why some welcome it)". In: (Nov. 2022). URL: https://www.billboard.com/pro/songs-getting-shorter-tiktok-streaming/.

[19] Elias Leight. "Pushing songs up the charts was a label job. then fans took over". In: (Aug. 2023). URL: https://www.billboard.com/pro/fans-pushing-songs-albums-up-charts-coordinated-efforts/.

[20] *LyricsGenius: a Python client for the Genius.com API — lyricsgenius documentation*. URL: https://lyricsgenius.readthedocs.io/en/master/.

[21] Kai Middlebrook and Kian Sheik. "Song Hit Prediction: Predicting Billboard Hits Using Spotify Data". In: *arXiv.org* (Sept. 2019). URL: https://arxiv.org/pdf/1908.08609.

[22] *pandas.DataFrame.combine_first — pandas 2.2.3 documentation*. URL: https://pandas.pydata.org/docs/reference/api/pandas.DataFrame.combine_first.html.

[23] *pandas.DataFrame.diff — pandas 2.2.3 documentation*. URL: https://pandas.pydata.org/docs/reference/api/pandas.DataFrame.diff.html.

[24] Harriman Samuel Saragih. "Predicting song popularity based on spotify's audio features: insights from the Indonesian streaming users". In: *Journal of Management Analytics* 10.4 (July 2023), pp. 693–709. DOI: 10.1080/23270012.2023.2239824. URL: https://doi.org/10.1080/23270012.2023.2239824.

[25] Mariangela Sciandra and Irene Carola Spera. "A model-based approach to Spotify data analysis: a Beta GLMM". In: *Journal of Applied Statistics* 49.1 (Aug. 2020), pp. 214–229. DOI: 10.1080/02664763.2020.1803810. URL: https://pmc.ncbi.nlm.nih.gov/articles/PMC9042099/.

[26] Danilo B. Seufitelli et al. "Hit song science: a comprehensive survey and research directions". In: *Journal of New Music Research* 52.1 (Jan. 2023), pp. 41–72. DOI: 10.1080/09298215.2023.2282999. URL: https://doi.org/10.1080/09298215.2023.2282999.

[27] Ali Shojaie and Emily B. Fox. "Granger Causality: a review and recent advances". In: *Annual Review of Statistics and Its Application* 9.1 (Nov. 2021), pp. 289–319. DOI: 10.1146/annurev-statistics-040120-010930. URL: https://doi.org/10.1146/annurev-statistics-040120-010930.

[28] André Spicer. "Taylor Swift's documentary shows why we should stop trying to be 'interesting'". In: (Feb. 2020). URL: https://www.theguardian.com/commentisfree/2020/feb/10/interesting-reinvention-taylor-swift-celebrities.

[29] Penn State. *11.2 Vector Autoregressive models VAR(p) models*. URL: https://online.stat.psu.edu/stat510/lesson/11/11.2.

[30] *statsmodels.tsa.seasonal.STL - statsmodels 0.14.4*. URL: https://www.statsmodels.org/stable/generated/statsmodels.tsa.seasonal.STL.html.

[31] *statsmodels.tsa.stattools.adfuller - statsmodels 0.15.0 (+522)*. URL: https://www.statsmodels.org/dev/generated/statsmodels.tsa.stattools.adfuller.html.

[32] Fernando Terroso-Saenz, Jesus Soto, and Andres Muñoz. "Evolution of global music trends: An exploratory and predictive approach based on Spotify data". In: *Entertainment Computing* 44 (Jan. 2023), p. 100536. DOI: `10.1016/j.entcom.2022.100536`. URL: `https://doi.org/10.1016/j.entcom.2022.100536`.

[33] *Web API Reference — Spotify for Developers*. URL: `https://developer.spotify.com/documentation/web-api/reference/get-audio-features`.

[34] *Welcome to Spotipy! — spotipy 2.0 documentation*. URL: `https://spotipy.readthedocs.io/en/2.24.0/`.