A Statistical Analysis of Taylor Swift's Music: Comparing Variables of Label Owned Records against Artist Owned Master Copies and Re-recordings

DSP 555 — Nellie Dawson, Ryan Soucy

Abstract

Consistently since her debut album release in 2006, Taylor Swift has made a profound impact on the music industry and beyond. A significant moment in Swift's career was her decision to leave her record label after her contract ended in 2018 in pursuit of gaining ownership and creative freedom of her music. The goal of this statistical analysis is to explore whether there is a fundamental difference between Swift's music in the factor groups designating before and after 2018. The raw data contains several metrics of Swift's discography and we analyze differences of these metrics between these two factor groups. We apply Hotelling's T-Square Test to compare group means, indicating that we reject the null hypothesis when all variable means between factor groups are not equal. We then perform K-Nearest Neighbors and Decision Tree classification methods and had moderate success at classifying Swift's songs to label or artist ownership.

Background

Taylor Swift's impact on pop culture is indisputable. From her concert attendees influencing seismic signals in Seattle (Sykes, 2023) to her strong sense of self-advocacy, it is undeniable that Swift is a powerful force in the world. In pursuit of ownership and creative freedom of her music, Swift was compelled to take on the extraordinary task of re-recording her first six albums over again to gain complete ownership of the new master copy versions. As she takes on this endeavor, we hope to investigate differences in her music before and after she was under an agreement with Big Machine Label Group. Has her music been the same since her debut in 2006, or has there been a fundamental shift in musical metrics of her songs since her release from her six-album contract with the Big Machine Label Group? This is the question we hope to answer through our analysis.

When she was fifteen years old, Swift signed a six-album agreement with a record label, but little did she know the self-advocacy work ahead of her. In 2014, she attempted negotiating a Spotify deal ultimately resulting in her own decision to remove her music from the streaming service for the following three years (Brandle, 2017). When Swift's contract with Big Machine Label Group was up in 2018, the newly acquired company did not allow her to use her previous songs in a documentary or perform them at an awards show (Tsioulcas, 2019). Not only was she determined to regain ownership of the master copies of her old albums, she was committed to controlling her music moving forward.

Taylor Swift simultaneously embarked on a mission to regain ownership of her previously recorded albums and recorded a brand new album, Lover, the first of which she completely owns. To regain ownership of her first six albums, she would re-record and release her previous six albums with (Taylor's Version) after each re-recorded album's title to distinguish that she owns the master copy versions. Since the 2021 re-release of her sophomore album, Fearless (Taylor's Version), Swift has released three total re-recorded albums. Additionally, she has released four other albums aside from these re-recordings since leaving her previous record label. With this new creative freedom and ownership of her music, we want to explore whether Taylor's style has changed since separating from Big Machine Label Group and in what ways. We hope to determine whether or not there is a fundamental difference in a variety of Spotify metrics between Taylor Swift's music released under the record label and now that she has complete ownership of the master recordings and if this difference is enough to perform classification methods of songs to artist or label ownership.

There are many sites and social media accounts dedicated to tracking Taylor Swift's every move. Within the fiercely loyal "Swiftie" community, some have taken their data skills to analyze Swift's song popularity in Top Billboard chart appearances (Wen & Scharfstein, 2022), and others have done reports on her album sales and her record-breaking debut week for her 2022 Midnights album (MJD, 2023). Trainor and Bankova's report *The Unstoppable Taylor Swift* explored features of her music over time by comparing each of her ten albums on just three variable metrics. This album comparison includes the Spotify metrics danceability, acousticness, and emotion (valence), three of the six variables we use in our research. Trainor and Bankova found consistencies across Swift's albums in danceability and valence measures but observed some distinct differences between albums in acousticness. Only the original versions of each album are included in their research analysis and no Taylor's Version rerecordings. In addition to using the deluxe edition of each album to increase observation size and be more representative of the population of her entire discography, we will be including both Swift's original recordings and her rerecordings in our study to attempt to identify if these features had a significant change since her original recordings of them.

In *An Exploratory Data Analysis of Taylor Swift's Music*, Aimi Wen and Ava Scharfstein explore the popularity of Swift's music and attempt to identify features and song lyrics that distinguished songs that made the top charts. Their biggest hurdle was the high dimensionality of songs on the top charts, so they tried to determine if there was any distinction in song features between whether the song performed very well or moderately well on the top charts. They used Logistic Regression and Decision Trees to attempt to classify songs to one of these factor levels but found limited success in using Spotify's API variables to predict chart popularity. With this consistently high popularity measure across Swift's career, we have removed popularity from the raw data in our analysis, and will not be relying on popularity as a response or predictor variable in our comparisons. We intend to compare the fundamental differences in the songs using these Spotify API variable measures.

We did not find any reports with this specific population comparison. This is the first time an artist has ever taken on such a large-scale re-recording project, so there aren't studies on the impact record labels have on music. We wanted to do something unique that hasn't been done yet, and with three more re-recordings on the way, we have the opportunity to revisit this future data to further our investigation and see how the answer to our research question will continue to evolve.

Data Description

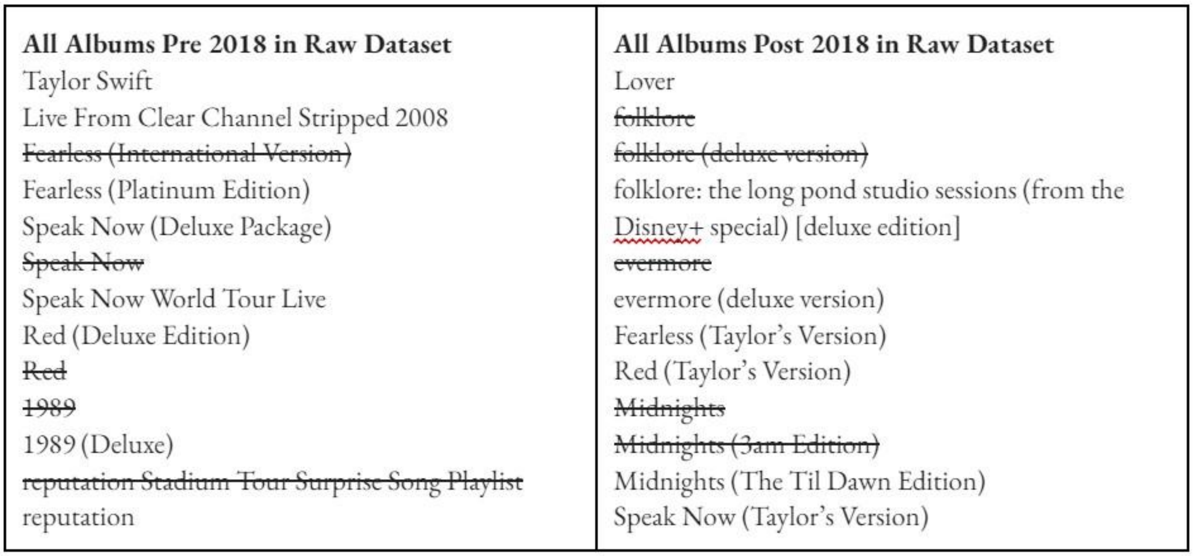
Our raw dataset "taylor_swift_spotify" specifically focuses on data from Spotify's API (Application Programming Interface) allowing app developers to use Spotify's song metrics in their programs. This raw data, made available on Kaggle by Jarred Priester, is a .csv file updated monthly. There are 486 rows of observations, songs in this case, and 18 columns. These columns contain the following variables: X, name, album, release_date, track_number, id, uri, acousticness, danceability, energy, instrumentalness, liveness, loudness, speechiness, tempo, valence, duration_ms, and popularity. Most numeric values are confidence measures from 0.0 to 1.0 while others are measurements such as BPM, decibels, and milliseconds for tempo, loudness, and duration, respectively. **Table 1** in the Appendix provides a detailed description of each variable and their measurement scales.

The raw dataset contains Swift's entire discography since her debut in 2006. Swift has a history of releasing an album and then subsequently releasing a deluxe version of that album that contains the exact same songs as the original release plus a few bonus tracks that didn't make the first cut. This means there are some exact duplicates of songs in this raw dataset and some similarly titled tracks that are, in fact, not duplicates. To remove exact song duplicates, we have chosen only to include the deluxe versions. By deleting the original release and keeping the deluxe, we avoid having subgroups of albums become unintentionally repeated data points in this analysis.

However, these removed subgroups within the deluxe albums are very different from the Taylor's Version of each album. These are the three albums Swift has re-recorded so far and they are viewed separately from the originally released albums. This is because these songs are not exact duplicates of the originals. These similarly titled tracks will actually provide the insight we are looking for when exploring the potential shift in musical metrics of Swift's first six albums compared to her true style when she had ownership and creative freedom over the re-recordings.

We have included all other albums released after leaving her previous label and the live albums within the raw dataset. Despite these being live versions of songs on other albums, they contain different values on Spotify's metrics due to their delivery method and, therefore, are representative of her music style before and after 2018 and could contribute to the study of exploring these differences. **Table 2** depicts the albums in the raw dataset, separated by whether they were released under her previous label or after. Albums with a ~~strikethrough~~ indicate it was removed to eliminate duplicates.

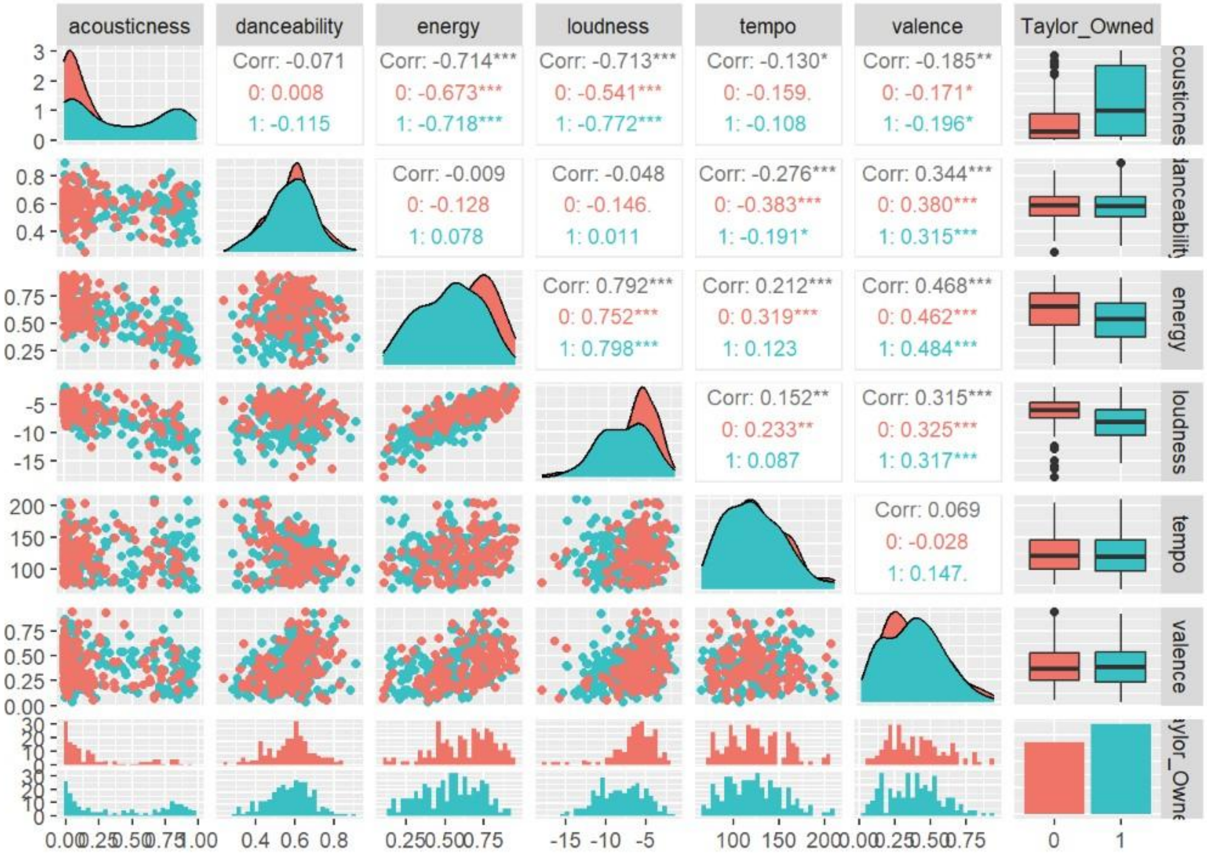
Table 2 — Albums from Raw and Selected Data



All Albums Pre 2018 in Raw Dataset	**All Albums Post 2018 in Raw Dataset**
Taylor Swift	Lover
Live From Clear Channel Stripped 2008	~~folklore~~
~~Fearless (International Version)~~	~~folklore (deluxe version)~~
Fearless (Platinum Edition)	folklore: the long pond studio sessions (from the Disney+ special) [deluxe edition]
Speak Now (Deluxe Package)	~~evermore~~
~~Speak Now~~	evermore (deluxe version)
Speak Now World Tour Live	Fearless (Taylor's Version)
Red (Deluxe Edition)	Red (Taylor's Version)
~~Red~~	~~Midnights~~
~~1989~~	~~Midnights (3am Edition)~~
1989 (Deluxe)	Midnights (The Til Dawn Edition)
~~reputation Stadium Tour Surprise Song Playlist~~	Speak Now (Taylor's Version)
reputation	

When considering the numerical variables available in the raw dataset, we looked for variables representative of the music in the dataset. Many of the variables from Spotify's metrics are a confidence measure from 0.0 to 1.0. There are no missing values in our dataset, but some variables have a large number of zero or almost zero values, making log and inverse transformations of these variables difficult. When looking at the distribution of the variables, it was clear that three of the 11 numeric variables do not represent music in this dataset: instrumentalness, liveness, and speechiness. Instrumentalness detects songs with limited to no vocals. Taylor's songwriting talents are present in her highly lyrical songs made clear in the summary statistics where the bottom 50% of the data all have the value of 0. Liveness detects the presence of a live audience which is only present on albums published before 2018, and removing this variable removes the bias of not having live albums released after 2018. Speechiness detects the presence of spoken word, and while most of her songs are melodic, this removes the potential of the test to be affected by three voice memos within the 1989 Deluxe Edition.

Figure 1 — GGplot (R) of Selected Variables



The six predictors used in this analysis are the Spotify metrics of acousticness, danceability, energy, loudness, tempo, and valence. As seen in **Figure 1** and confirmed by the Mardia test**,** danceability presents as univariately normal in both groups, and valence has a univariate normal distribution in just the artist-owned factor grouping. When comparing the distribution plots and scatter plot matrices of the variables, there are some overlaps between groups, but there are visual indications that there may be some differences. Variables that show the most distinction between artist and label-owned factor groupings include acousticness, energy, and loudness. There is some collinearity between variables, which aligns with the impact on each other in music. Acousticness has a strong negative correlation with both loudness and energy, while loudness and energy have a strong positive correlation. Valence, indicating happy or sad emotion, has a moderate correlation with danceability, energy, and loudness.

Methods

Through our analysis, we consider differences in Swift's music between her label-owned and artist-owned albums. Our data frame will include an index variable, the six Spotify metric predictors, and a created factor variable distinguishing label and artist ownership. The created factor variable holds a 0 value if the album release_date is before 2018 and a value of 1 if the album release_date is after 2018 distinguishing separation between label-owned and artist-owned albums. This factor splits the data into the comparison groups at the basis of our investigation in determining if Taylor's re-recordings and new music have changed in these musical metrics since leaving the label.

Through our exploration of data normality, the Mardia test indicated that the data is rarely univariately normal and, therefore, not multivariately normal. After seeing some visual group differences in the graphical data representations, we compared the factor group means for each variable using Hotelling's T-Square test. Hotelling's T-Squared test takes two sample groups and assesses each variable's mean. This test does make some assumptions about the data's independence, distributions, and covariance matrices. Our data meets all of these assumptions except for multivariate normality.

Because we do not meet the strict assumptions of parametric methods, our results would be very unstable, making those tests unreliable for our data. Therefore, we could not use the parametric classification methods of Linear and Quadratic Discriminant Analysis. We also could not use Logistic Regression confidently due to the high level of collinearity in the data. This narrows our focus to the nonparametric methods of K-Nearest Neighbors and Decision Trees to test the ability to classify data as label or artist-owned based on the selected predictor variables. Both the KNN and Decision Tree methods include building a model on a training set and finding a test error with a test set previously made through 80/20 percent probabilities. We also used 5-fold cross-validation for KNN and cross-validation for pruning the Decision Tree. We do all of this analysis in hopes of answering whether there is a significant difference in selected Spotify metric variables between label and artist-owned master copies, and if this difference will be enough to perform classification methods to classify songs to label or artist ownership.

K-Nearest Neighbors assesses nearby data points to determine which classification the song observation belongs to. The data's decision boundary is clearly not linear from the graphical representations, and we hope this might be a successful method of classifying whether songs are label or artist-owned. Despite the lack of assumptions, KNN is limited by the curse of dimensionality when there are too many predictor variables. It was beneficial for the method that we paired down the variables from the raw data set, from 11 numeric variables to six. Although nonparametric methods make no assumptions about the data, these methods can struggle to perform well when the number of predictors is too large and benefit from a large number of observations. Taylor Swift's discography allows us to perform this analysis on 306 song observations.

Decision Trees assess variables by splitting the variables at nodes based on the magnitude of the variables in the dataset. Each split creates regions that attempt to mimic the decision-making process of classification of data points. As a data point follows the tree, it is classified into regions of the artist and label-owned factor variable. Classification trees are especially useful because they are highly interpretable and are thought to mimic human decision-making by providing important variables with which to classify the data. The biggest limitation of Decision Trees is that some data is just not meant to be separated into node regions accurately, specifically linear data. However, our data is largely nonlinear, so it appears to comply with the use of Decision Tree Classification Methods.

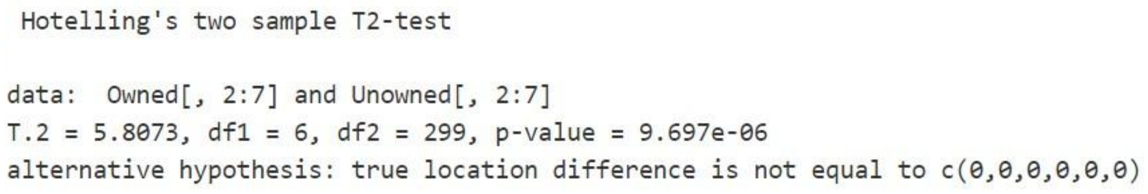
Data Analysis and Main Results

Hotelling's T-Square Test

As we prepare for Hotelling's T-Square test, we consider the assumptions of independence, distributions, and common covariance matrices. While the data is not multivariate normal, we previously removed variables from the raw data that were the most skewed. This method assumes common variance-covariance matrices, and the size and magnitude of the trends within the covariance matrices across factor groups are relatively comparable. The only variable that very clearly did not meet this criterion was duration, but this is a variable we previously removed from the raw data for this analysis. Each variable is a measurement of a Spotify metric, and each song's metric is independent of the other songs. By removing all of the exact duplicate songs, we ensured that every value is independent of the other and have no sub-populations with mean vectors. Thus, our data meets all assumptions except for multivariate normality.

In terms of hypothesis testing, a null hypothesis would state that all variables have equal means when comparing label and artist-owned songs. The alternative rejects the null hypothesis when at least one variable has a substantial difference when comparing groupings of label and artist-owned master copies. If the difference for each variable is zero, we would accept the null hypothesis that these group means are equal. If the difference for at least one variable is not zero, we will reject the null hypothesis and turn to the alternative hypothesis that the "difference in means is unequal to vector c(0,0,0,0,0,0)" indicating a vector of six zeros for the six variables. As shown in **Figure 2,** Hotelling's T-Squared test does indicate with a p-value of almost zero (9.697e-06) that at least one mean is unequal when comparing the groupings made by the artist and label ownership factor variable. Once Hotelling's T-Squared test indicates that there may be a difference between the group means, we move to classification methods to attempt to classify data points into these ownership factor groups.

Figure 2 — Results (R) Hotelling's T-Squared test



```
 Hotelling's two sample T2-test

data:  Owned[, 2:7] and Unowned[, 2:7]
T.2 = 5.8073, df1 = 6, df2 = 299, p-value = 9.697e-06
alternative hypothesis: true location difference is not equal to c(0,0,0,0,0,0)
```

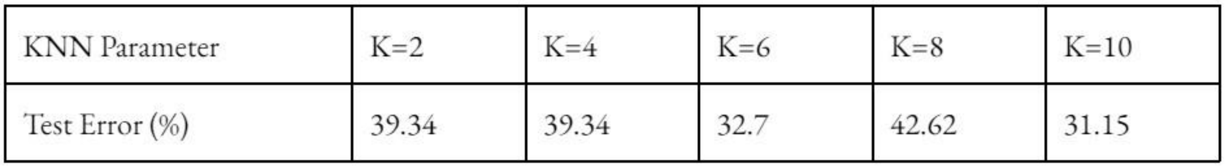
K-Nearest Neighbors

Our goal with KNN classification is to see if the method can successfully classify any song data point in the test set to the correct artist or label ownership factor level. We chose KNN because of its nonparametric nature, meaning there are no assumptions the data needs to meet before attempting the method, and the only parameter is determining the number of K nearest neighbors to choose.

Although there are no assumptions needed to be met, there is still the curse of dimensionality which stipulates that a large number of variables in the KNN model will weaken the stability of the model.

For the KNN classification approach, we tested multiple parameter sizes in conjunction with cross-validation to achieve the best estimate of a test error with our dataset with the most optimal parameter selection. Test errors on each parameter ranged from 31.15% to 42.62%. The results of this parameter testing are shown in **Table 3**.

Table 3 — Testing KNN Parameter Size



KNN Parameter	K=2	K=4	K=6	K=8	K=10
Test Error (%)	39.34	39.34	32.7	42.62	31.15

By computing a variety of testing on the trained models, we assessed K values ranging from 2 to 10 to see which would give us the best test error. We risk overfitting when using the K=1 nearest neighbor, so we have excluded it from this analysis. The KNN error rates fluctuated at various levels of K in this range. To avoid underfitting with too large of a K value, we determined the best test error rate in the range of 2 to 10 is six which gives us an error rate of 32.79%.

We then ran a 5-fold cross-validation procedure to test the effectiveness of this size parameter. The test errors ranged from 29.5% to 37.7% with an average of 32.4%. This misclassification is comparable to the mean we saw above which solidifies K=6 as an optimal parameter choice. A limitation of the K-Nearest Neighbor approach is that it does not indicate influential predictors, so we turn to Decision trees to see which variables are most important in this classification process.

Decision Trees

Next, we used a series of classification tree models to predict the same artist and label ownership factor variable. The first classification tree model, before pruning, includes 22 terminal nodes. This full tree has a misclassification rate of about 22.86% and only uses 5 of the variables to create the tree: loudness, energy, acousticness, danceability, and valence. We pruned this tree using cross-validation, and **Figure 3** shows that the smallest number of misclassifications is exhibited by trees of size 16 and above. This pruned tree with the same misclassification rate as the full tree, is shown in **Figure 4**, which has a test error of about 39.34%.

Figure 3 — Cross Validation Pruning

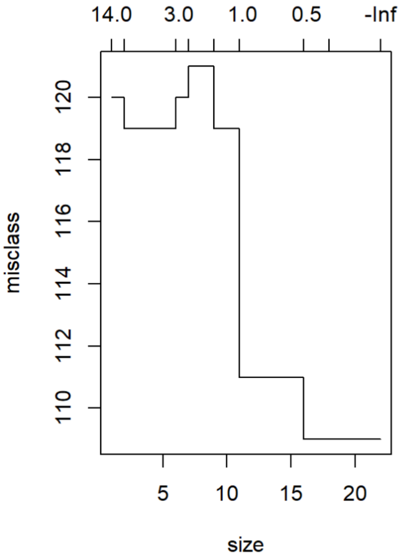
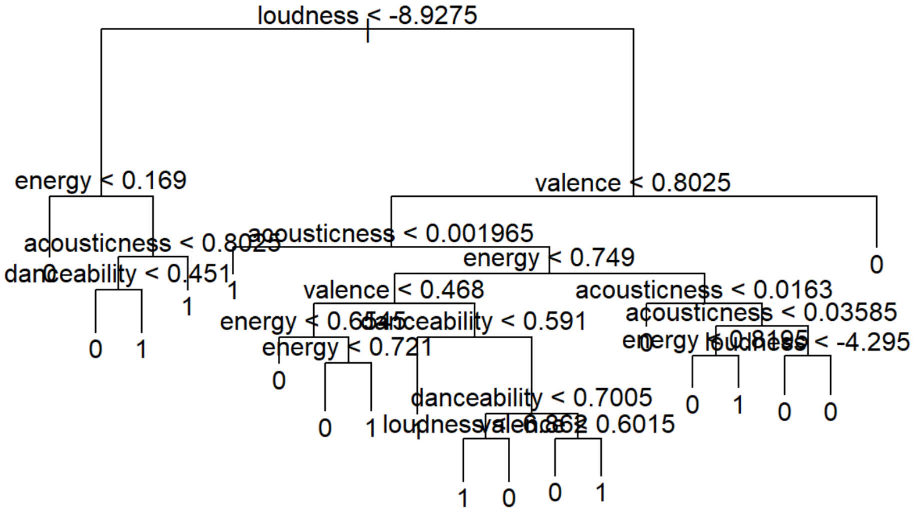


Figure 4 — Pruned Tree



The first and most impactful node is a split using the loudness variable at -8.9275, meaning that songs that are quieter than this value follow the branch left, and songs louder than this value will follow the branch to the right. It is important to note that the minimum value for loudness is recorded at -17.93, and the maximum value is -1.91. The bottom 25% of the loudness values are between -17.93 and -9.20. This first node indicates that somewhere between 25 and 50 percent of the data is pulled to the left side of the node while the rest continues right.

A node of interest in the first three is the valence node on the right side of the tree after the loudness node. Valence is the measurement of emotion in the song. If the valence is close to 0, the song is considered sad, while a valence value close to 1, indicates that the song is happy. The valence node splits at a value of 0.8025 and ends in a 0. This loosely means that every song that is happiest in this loud region of the data is a song under label ownership. Though outside the scope of this analysis, this begs the question of if the label influenced Swift's music to be either louder and/or happier.

The next important node is the energy node which splits at 0.169 and ends in a leaf of 0. Songs that are less loud and have lower energy are also not owned by Taylor. This appears to be the opposite of the loud and happy music observed previously, so it is equally important to note and question. Beyond these three splits, the data gets more difficult to interpret.

We then performed Random Forest on the data. We chose to use Random Forest over Bagging because the randomized trees produced through Bagging are likely to have a high correlation to each other. If one variable, in this case, loudness, is stronger than the others, it will likely be at the top of the majority of the Bagged trees and cause the trees to look similar to each other, therefore being highly correlated. Finding a mean of correlated data does little to decrease the variance in the data. Random Forest helps alleviate this overfitting by considering a limited number of available predictors. The predictors for each of the random forest trees are different for each tree. This results in an accurate interpretation of the accuracy of the model. Classification trees are prone to overfitting the data, and Random Forests help alleviate this. The Random Forest model had a mean test error of about 37.70 %.

Conclusions and Research Directions

In conclusion, despite the overlap in the factor groupings, Hotelling's T-Square test finds a distinction in these group means. It is clear through graphical analysis and classification methods that some variables were more present in her songs produced by the record label. We cannot make any claims about why the music may be different under label management, just that the groups appear to have differences. These differences are enough to classify the song observations moderately successfully on the ownership variable. The most successful K-Nearest Neighbor parameter provided cross-validated test errors of about 32%, while the most successful Decision tree method, random forest, gave a test error rate of approximately 37.7%.

At the time of writing this report, Taylor Swift is on the precipice of releasing a 4th Taylor's Version re-recording, 1989 (Taylor's Version). Although unannounced, there are two final albums needed to fulfill her mission of gaining ownership of all six master re-recordings. With these new albums on the horizon, we have multiple opportunities to further analyze the data. One possibility is to test the currently trained models on the new albums to assess their classification abilities. Additionally, we could introduce these new albums to the training data and observe the impact on error rates and decision trees.

SOURCES

Brandle, L. (2017, June 9). Taylor Swift's entire catalog is now on Spotify & Other Streaming Services: Go listen. Billboard. https://www.billboard.com/music/music-news/taylor-swifts-entire-catalog-spotify-streaming-7825588/

MJD. (2023, August 9). *Taylor Swift Albums and songs sales (updated daily)*. ChartMasters. https://chartmasters.org/taylor-swift-albums-and-songs-sales/

Sykes, J., & Rosenbloom, A. (2023, July 28). Taylor Swift Fans "shake it off," causing record-breaking seismic activity during Seattle shows. CNN. https://www.cnn.com/2023/07/27/entertainment/taylor-swift-seismic-activity/index.html

Trainor, C., & Bankova, D. (2023, July 29). *The unstoppable pop of Taylor Swift*. Reuters. https://www.reuters.com/graphics/MUSIC-TAYLORSWIFT/SPOTIFY/dwpkarywqpm/

Tsioulcas, A. (2019, November 15). *Taylor Swift's TV drama: Pop star claims former label won't let her perform hits*. NPR. https://www.npr.org/2019/11/15/779692984/taylor-swifts-tv-drama-pop-star-claims-former-label-wont-let-her-perform-hits

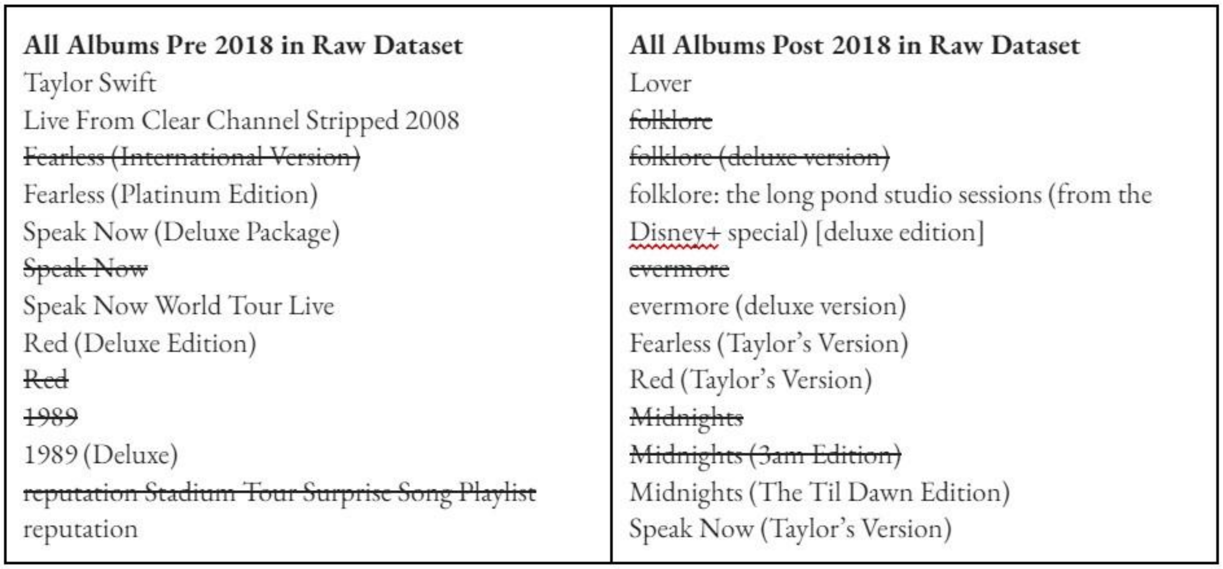
Wen, A., & Scharfstein, A. (2022, December 9). *Exploring Taylor Swift's music*. An Exploratory Data Analysis of Taylor Swift's Music. https://medium.com/@aimiwen33/exploring-taylor-swifts-music-2bce11a7aab2

APPENDIX

Table 1 — Variable Descriptions and Scales

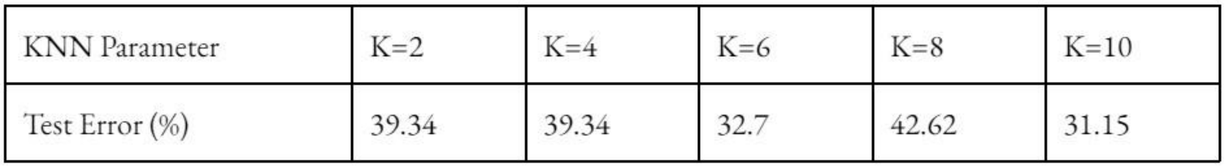
Measure	Scale	*Definitions from CRAN - R Documentation on Spotify API Data*
Acousticness	0.0 – 1.0	*A confidence measure from 0.0 to 1.0 of whether the track is acoustic. 1.0 represents high confidence the track is acoustic.*
Danceability	0.0 – 1.0	*Danceability describes how suitable a track is for 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.*
Energy	0.0 – 1.0	*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. For example, death metal has high energy, while a Bach prelude scores low on the scale. Perceptual features contributing to this attribute include dynamic range, perceived loudness, timbre, onset rate, and general entropy.*
Instrumentalness	0.0 – 1.0	*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 1.0, 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 1.0.*
Liveness	0.0 – 1.0	*Detects the presence of an audience in the recording. Higher liveness values represent an increased probability that the track was performed live. A value above 0.8 provides strong likelihood that the track is live.*
Loudness	-60 – 0	*The overall loudness of a track in decibels (dB). Loudness values are averaged across the entire track and are useful for comparing relative loudness of tracks. Loudness is the quality of a sound that is the primary psychological correlate of physical strength (amplitude). Values typical range between -60 and 0 db.*
Speechiness	0.0 – 1.0	*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. Values above 0.66 describe tracks that are probably made entirely of spoken words. Values between 0.33 and 0.66 describe tracks that may contain both music and speech, either in sections or layered, including such cases as rap music. Values below 0.33 most likely represent music and other non-speech-like tracks.*
Tempo	BPM 60 – 220	*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	0.0 – 1.0	*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).*
Popularity	0 – 100	*The frequency that a track has been played (Spotify API)*

Table 2 — Albums from Raw and Selected Data



All Albums Pre 2018 in Raw Dataset	**All Albums Post 2018 in Raw Dataset**
Taylor Swift	Lover
Live From Clear Channel Stripped 2008	~~folklore~~
~~Fearless (International Version)~~	~~folklore (deluxe version)~~
Fearless (Platinum Edition)	folklore: the long pond studio sessions (from the Disney+ special) [deluxe edition]
Speak Now (Deluxe Package)	~~evermore~~
~~Speak Now~~	evermore (deluxe version)
Speak Now World Tour Live	Fearless (Taylor's Version)
Red (Deluxe Edition)	Red (Taylor's Version)
~~Red~~	~~Midnights~~
~~1989~~	~~Midnights (3am Edition)~~
1989 (Deluxe)	Midnights (The Til Dawn Edition)
~~reputation Stadium Tour Surprise Song Playlist~~	Speak Now (Taylor's Version)
reputation	

Table 3 — Testing KNN Parameter Size



KNN Parameter	K=2	K=4	K=6	K=8	K=10
Test Error (%)	39.34	39.34	32.7	42.62	31.15

Figure 1 – 4 Source — R code attached

Final Project - Taylor Swift Analysis

Nellie Dawson, Ryan Soucy

DPS 555

Raw and Selected Data

```
#Raw Data
RawData <- read.csv("taylor_swift_spotify.csv")
names(RawData)
```

```
##  [1] "X"                "name"             "album"            "release_date"
##  [5] "track_number"     "id"               "uri"              "acousticness"
##  [9] "danceability"     "energy"           "instrumentalness" "liveness"
## [13] "loudness"         "speechiness"      "tempo"            "valence"
## [17] "popularity"       "duration_ms"
```

18 CATEGORICAL AND QUANTITATIVE VARIABLES

```
#Creating New Factor Variable
RawData$Taylor_Owned <- 1*(RawData$release_date > 2018)
RawData$Taylor_Owned = as.factor(RawData$Taylor_Owned)
```

```
#Selecting Albums (removing duplicate songs)
UnownedAlbums <- filter(RawData, album %in% c("Taylor Swift", "Live From Clear Channel Stripped
2008", "Fearless (Platinum Edition)" , "Speak Now (Deluxe Package)", "Speak Now World Tour Liv
e", "Red (Deluxe Edition)" , "1989 (Deluxe)", "reputation"))

OwnedAlbums  <- filter(RawData, album %in% c("Lover","folklore: the long pond studio sessions (f
rom the Disney+ special) [deluxe edition]", "evermore (deluxe version)", "Fearless (Taylor's Ver
sion)" , "Red (Taylor's Version)", "Speak Now (Taylor's Version)", "Midnights (The Til Dawn Edit
ion)"))

SelectAlbums <- rbind(OwnedAlbums, UnownedAlbums)
```

```
#Selected Variables
SelectData0 <- SelectAlbums[,8:19]
Index <- SelectAlbums[,1]

#Combine Index and Selected Albums
SelectData <- cbind(Index,SelectData0)

names(SelectData)
```

```
##  [1] "Index"            "acousticness"     "danceability"     "energy"          
##  [5] "instrumentalness" "liveness"         "loudness"         "speechiness"     
##  [9] "tempo"            "valence"          "popularity"       "duration_ms"     
## [13] "Taylor_Owned"
```

NUMERIC VARIABLES: acousticness, danceability, energy, instrumentalness, liveness, loudness, speechiness, tempo, valence, popularity, and duration.

```
Summaries <- sapply(SelectData[,2:12], summary)
Summaries
```

```
##         acousticness danceability    energy instrumentalness  liveness
## Min.       0.0001910    0.2430000 0.1180000      0.000000000 0.0335000
## 1st Qu.    0.0329250    0.5052500 0.4447500      0.000000000 0.0962000
## Median     0.1630000    0.5880000 0.5865000      0.000000000 0.1165000
## Mean       0.3272365    0.5763889 0.5728072      0.002612165 0.1783196
## 3rd Qu.    0.6685000    0.6485000 0.7315000      0.000026650 0.1770000
## Max.       0.9710000    0.8970000 0.9440000      0.328000000 0.9310000
##           loudness speechiness     tempo   valence popularity duration_ms
## Min.    -17.932000  0.02310000  68.09700 0.0382000    0.00000     83253.0
## 1st Qu.  -9.204750  0.02960000  96.98825 0.2395000   53.00000    214333.2
## Median   -6.732000  0.03605000 120.02050 0.3830000   66.00000    237302.0
## Mean     -7.346255  0.05763105 123.65716 0.3963824   63.60131    244048.0
## 3rd Qu.  -5.163000  0.05387500 145.86450 0.5327500   74.00000    263955.8
## Max.     -1.909000  0.91200000 208.91800 0.9420000   98.00000    613026.0
```

SUMMARY: Variables whose mean higher than median (Skewed Right): ACOUSTICNESS, INSTRUMENTALNESS (Bottom 50% of data is all 0), SPEECHINESS. Other variables have mean and median values reasonably close in range.

Covariance Matrices by Group

```
#Owed/Unowned Subgroups
Own <- subset(SelectData, Taylor_Owned == 1)
Unown <- subset(SelectData, Taylor_Owned == 0)

#Owned
CovarianceOwned <- round(cov(Own[,2:12]),5)
CovarianceOwned
```

```
##                   acousticness danceability    energy instrumentalness  liveness
## acousticness          0.12704     -0.00470  -0.04846          0.00151  -0.00498
## danceability         -0.00470      0.01316   0.00170         -0.00024  -0.00022
## energy               -0.04846      0.00170   0.03585         -0.00033   0.00336
## instrumentalness      0.00151     -0.00024  -0.00033          0.00097  -0.00009
## liveness             -0.00498     -0.00022   0.00336         -0.00009   0.00845
## loudness             -0.80795      0.00358   0.44400         -0.01696   0.05043
## speechiness          -0.00068      0.00147  -0.00048         -0.00005  -0.00017
## tempo                -1.20412     -0.68206   0.72517          0.00625  -0.18773
## valence              -0.01344      0.00695   0.01766         -0.00039  -0.00212
## popularity           -1.87843      0.11465   0.91426         -0.01035   0.01726
## duration_ms        -637.99061  -1582.69814 -54.52761        -80.70059 -30.59685
##                     loudness speechiness       tempo     valence  popularity
## acousticness        -0.80795    -0.00068    -1.20412    -0.01344    -1.87843
## danceability         0.00358     0.00147    -0.68206     0.00695     0.11465
## energy               0.44400    -0.00048     0.72517     0.01766     0.91426
## instrumentalness    -0.01696    -0.00005     0.00625    -0.00039    -0.01035
## liveness             0.05043    -0.00017    -0.18773    -0.00212     0.01726
## loudness             8.63044    -0.02445     7.97927     0.17949    16.63252
## speechiness         -0.02445     0.00316    -0.04359     0.00112     0.04568
## tempo                7.97927    -0.04359   970.00693     0.88318    16.54403
## valence              0.17949     0.00112     0.88318     0.03708     0.42363
## popularity          16.63252     0.04568    16.54403     0.42363    94.63961
## duration_ms      20571.70315  -938.84319 10198.08865 -2606.15790 73348.14159
##                    duration_ms
## acousticness     -6.379906e+02
## danceability     -1.582698e+03
## energy           -5.452761e+01
## instrumentalness -8.070059e+01
## liveness         -3.059685e+01
## loudness          2.057170e+04
## speechiness      -9.388432e+02
## tempo             1.019809e+04
## valence          -2.606158e+03
## popularity        7.334814e+04
## duration_ms       2.546568e+09
```

```
#Unowned
CovarianceUnowned <- round(cov(Unown[,2:12]),5)
CovarianceUnowned
```

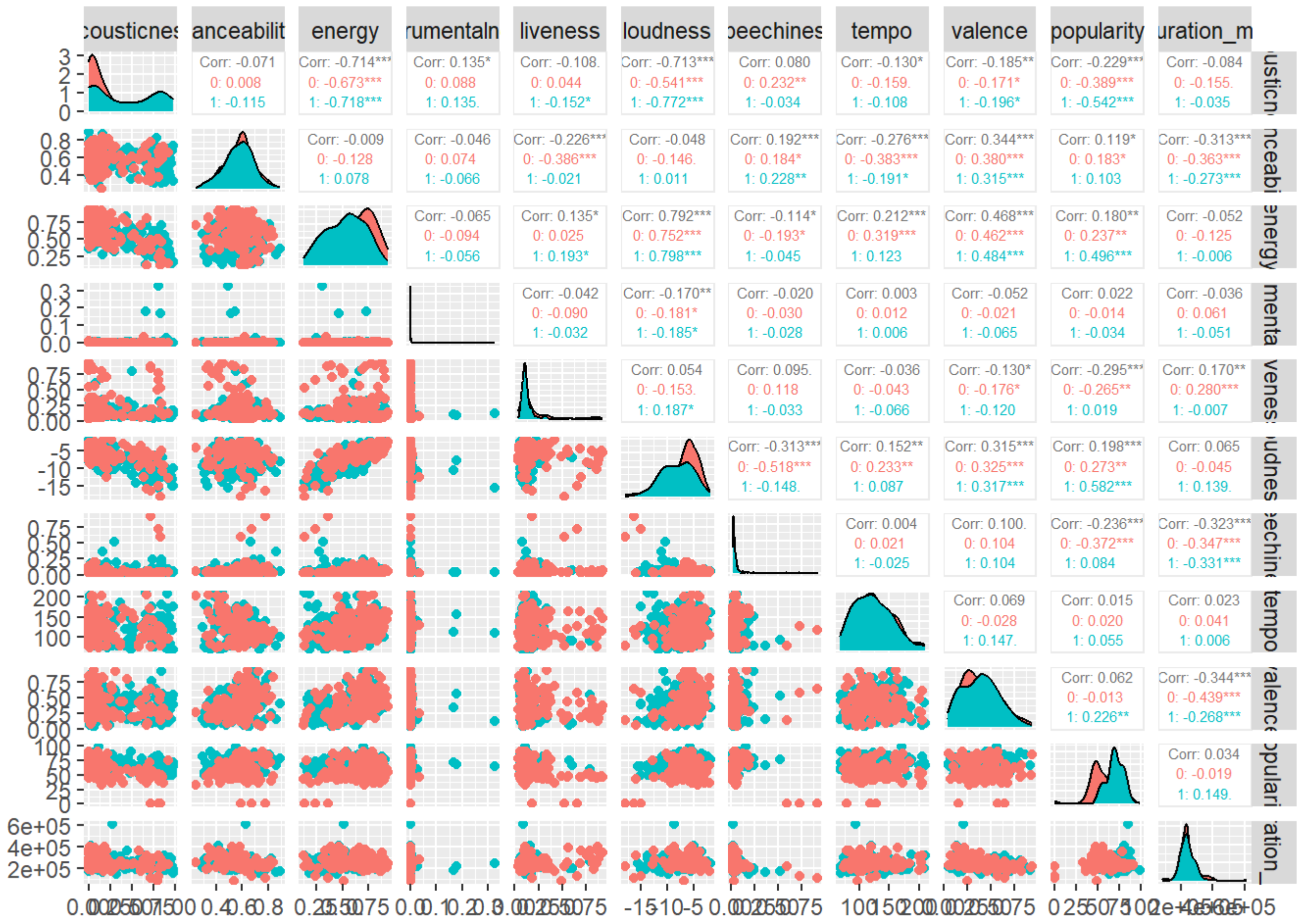
```
##                  acousticness danceability      energy instrumentalness
## acousticness          0.07189      0.00024    -0.03442          0.00007
## danceability          0.00024      0.01311    -0.00279          0.00003
## energy               -0.03442     -0.00279     0.03641         -0.00005
## instrumentalness      0.00007      0.00003    -0.00005          0.00001
## liveness              0.00267     -0.00999     0.00108         -0.00006
## loudness             -0.39054     -0.04499     0.38655         -0.00146
## speechiness           0.00677      0.00230    -0.00402         -0.00001
## tempo                -1.34280     -1.37723     1.91215          0.00110
## valence              -0.00894      0.00850     0.01724         -0.00001
## popularity           -1.67440      0.33529     0.72572         -0.00067
## duration_ms       -2143.96305  -2142.07336 -1229.11786          9.43373
##                    liveness    loudness speechiness       tempo     valence
## acousticness        0.00267    -0.39054     0.00677    -1.34280    -0.00894
## danceability       -0.00999    -0.04499     0.00230    -1.37723     0.00850
## energy              0.00108     0.38655    -0.00402     1.91215     0.01724
## instrumentalness   -0.00006    -0.00146    -0.00001     0.00110    -0.00001
## liveness            0.05114    -0.09328     0.00291    -0.30472    -0.00777
## loudness           -0.09328     7.26172    -0.15226    19.72907     0.17142
## speechiness         0.00291    -0.15226     0.01190     0.07118     0.00222
## tempo              -0.30472    19.72907     0.07118   985.98288    -0.17311
## valence            -0.00777     0.17142     0.00222    -0.17311     0.03820
## popularity         -0.96132    11.79093    -0.65131    10.15104    -0.03965
## duration_ms      3266.81221 -6201.11933 -1950.62038 67044.70247 -4421.42280
##                    popularity   duration_ms
## acousticness         -1.67440 -2.143963e+03
## danceability          0.33529 -2.142073e+03
## energy                0.72572 -1.229118e+03
## instrumentalness     -0.00067  9.433730e+00
## liveness             -0.96132  3.266812e+03
## loudness             11.79093 -6.201119e+03
## speechiness          -0.65131 -1.950620e+03
## tempo                10.15104  6.704470e+04
## valence              -0.03965 -4.421423e+03
## popularity          257.09608 -1.540303e+04
## duration_ms      -15403.02865  2.659370e+09
```

COVARIANCE: Loudness, Tempo, and Duration have the most variance. Aside from duration, other variables seem to have somewhat similar trends in direction and magnitude.

Graphical Data Summaries by Group

(Distribution, Scatter Plot, Correlation)

```
#Using ggpairs to create scatter plot correlation matrices colored by Taylor Owned
ggpairs(SelectData[,2:12], aes(color = SelectData$Taylor_Owned), upper = list(continuous = wrap
("cor", size=2)))
```



RED: UNOWNED (Pre 2018)

BLUE: OWNED (Post 2018)

DISTRIBUTIONS: Acousticness shows an interesting trend within and when comparing groupings - Almost bimodal in blue. Instrumentalness, Liveness, Speechiness seem to have heavy right skew

SPLOM: Some distinction between colors (energy, energy, loudness, valence, acousticness)

CORRELATIONS: High collinearity. Similar trends when comparing groupings

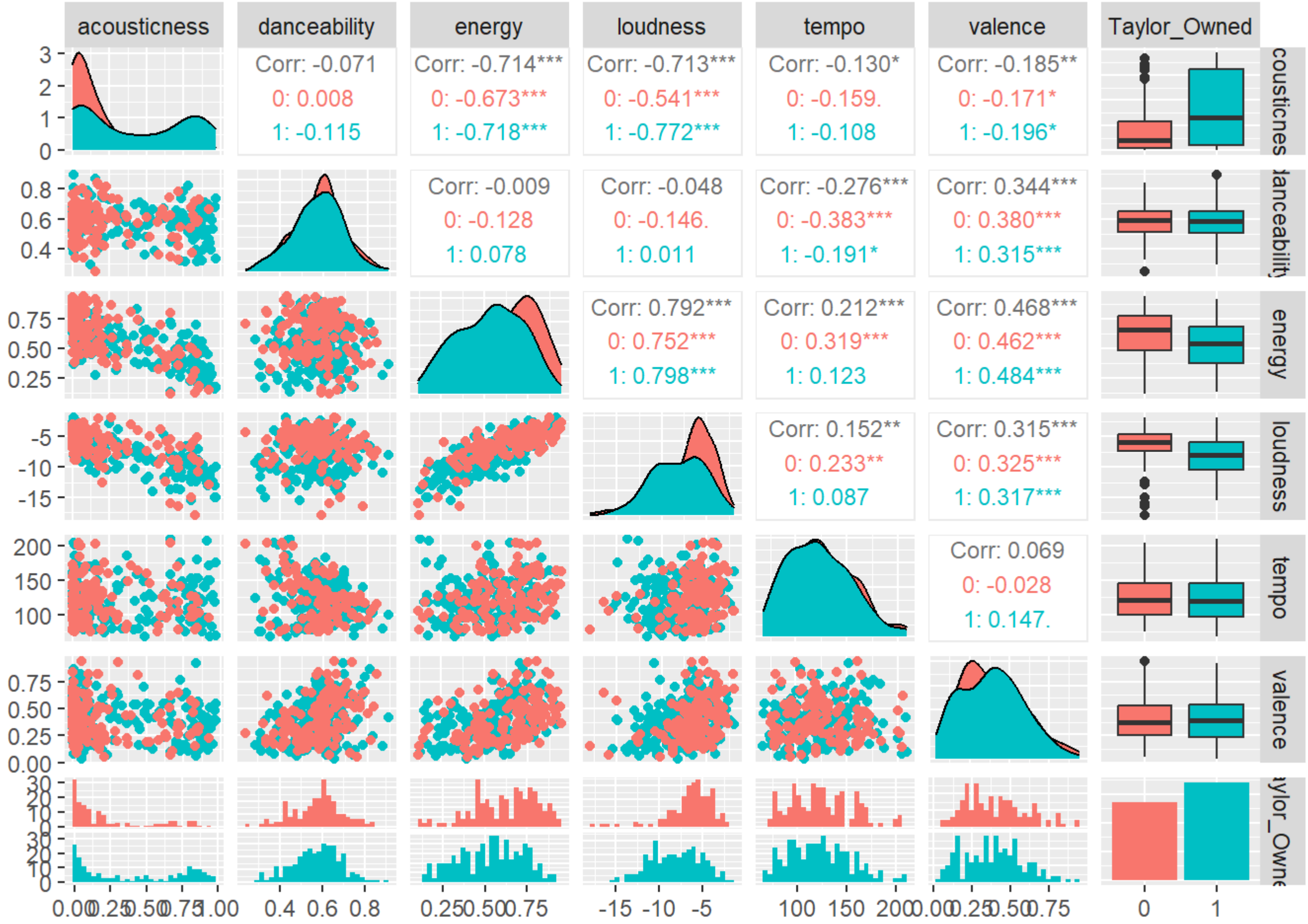
Selected Variables - New Ggplot

```
SelectData <- SelectData[ ,!names(SelectData) %in% c("speechiness","liveness","instrumentalnes
s","popularity","duration_ms")]
names(SelectData)
```

```
## [1] "Index"        "acousticness" "danceability" "energy"       "loudness"
## [6] "tempo"        "valence"      "Taylor_Owned"
```

```
ggpairs(SelectData[,2:8], aes(color = SelectData$Taylor_Owned), upper = list(continuous = wrap
("cor", size=3)))
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



Checking Normality

```
#Multivariate Normality
set.seed(3)
by(SelectData[,2:7], SelectData$Taylor_Owned, mvn)
```

```
## SelectData$Taylor_Owned: 0
## $multivariateNormality
##             Test       HZ p value MVN
## 1 Henze-Zirkler 1.511835       0  NO
## 
## $univariateNormality
##                Test     Variable Statistic   p value Normality
## 1 Anderson-Darling acousticness   12.7727  <0.001      NO    
## 2 Anderson-Darling danceability    0.7278  0.0564      YES   
## 3 Anderson-Darling    energy       1.3100   0.002      NO    
## 4 Anderson-Darling   loudness      3.6935  <0.001      NO    
## 5 Anderson-Darling    tempo        1.0791  0.0076      NO    
## 6 Anderson-Darling   valence       1.3482  0.0016      NO    
## 
## $Descriptives
##                n        Mean    Std.Dev    Median        Min     Max     25th
## acousticness 136   0.2186331  0.2681303   0.09855   0.000197   0.921  0.01845
## danceability 136   0.5786838  0.1144865   0.59200   0.243000   0.843  0.51250
## energy       136   0.6242941  0.1908143   0.65950   0.118000   0.944  0.48175
## loudness     136  -6.3806397  2.6947584  -5.88800 -17.932000  -1.953 -7.36000
## tempo        136 124.7653971 31.4003643 120.52700  74.900000 204.489 99.70425
## valence      136   0.4031559  0.1954458   0.37400   0.049900   0.942  0.25125
##                   75th       Skew    Kurtosis
## acousticness   0.29150  1.2472922  0.11589068
## danceability   0.64700 -0.2259193 -0.05908314
## energy         0.77700 -0.5309268 -0.33033410
## loudness      -4.67775 -1.6603554  4.03188345
## tempo        145.86925  0.4564514 -0.41431719
## valence        0.52400  0.5517683 -0.33003199
## 
## ------------------------------------------------------------
## SelectData$Taylor_Owned: 1
## $multivariateNormality
##             Test       HZ      p value MVN
## 1 Henze-Zirkler 1.261969 1.457409e-10  NO
## 
## $univariateNormality
##                Test     Variable Statistic   p value Normality
## 1 Anderson-Darling acousticness    9.4402  <0.001      NO    
## 2 Anderson-Darling danceability    0.3796  0.4006      YES   
## 3 Anderson-Darling    energy       0.9120  0.0198      NO    
## 4 Anderson-Darling   loudness      0.7602  0.0471      NO    
## 5 Anderson-Darling    tempo        0.9440  0.0165      NO    
## 6 Anderson-Darling   valence       0.7285  0.0565      YES   
## 
## $Descriptives
##                n        Mean    Std.Dev   Median        Min     Max      25th
## acousticness 170   0.4141192  0.3564243   0.3260   0.000191   0.971   0.05050
## danceability 170   0.5745529  0.1147185   0.5840   0.292000   0.897   0.50525
## energy       170   0.5316176  0.1893337   0.5455   0.131000   0.915   0.37625
## loudness     170  -8.1187471  2.9377608  -7.9770 -15.489000  -1.909 -10.48400
## tempo        170 122.7705706 31.1449343 119.7730  68.097000 208.918  96.31100
## valence        170   0.3909635  0.1925733   0.3950   0.038200   0.920   0.23375
##                    75th       Skew    Kurtosis
## acousticness   0.80325  0.1901600 -1.65244201
## danceability   0.65075 -0.1214528 -0.01516759
## energy         0.68575 -0.1415555 -0.90152873
## loudness      -5.85125 -0.1774320 -0.68790043
## tempo        145.37950  0.4539709 -0.32662161
## valence        0.53275  0.2499692 -0.60199622
```

UNIVARIATE NORMALITY: Danceability, Valence MULTIVARIATE NORMALITY: Not multivariate normal

```
#Owned/Unowned Subgroups
Owned <- subset(SelectData, Taylor_Owned == 1)
Unowned <- subset(SelectData, Taylor_Owned == 0)
```

New Covariance Matrices by Group

```
#Owned
CovarianceOwned <- round(cov(Owned[,2:7]),5)
CovarianceOwned
```

```
##              acousticness danceability   energy loudness     tempo  valence
## acousticness      0.12704     -0.00470 -0.04846 -0.80795  -1.20412 -0.01344
## danceability     -0.00470      0.01316  0.00170  0.00358  -0.68206  0.00695
## energy           -0.04846      0.00170  0.03585  0.44400   0.72517  0.01766
## loudness         -0.80795      0.00358  0.44400  8.63044   7.97927  0.17949
## tempo            -1.20412     -0.68206  0.72517  7.97927 970.00693  0.88318
## valence          -0.01344      0.00695  0.01766  0.17949   0.88318  0.03708
```

```
#Unowned
CovarianceUnowned <- round(cov(Unowned[,2:7]),5)
CovarianceUnowned
```

```
##              acousticness danceability   energy loudness     tempo  valence
## acousticness      0.07189      0.00024 -0.03442 -0.39054  -1.34280 -0.00894
## danceability      0.00024      0.01311 -0.00279 -0.04499  -1.37723  0.00850
## energy           -0.03442     -0.00279  0.03641  0.38655   1.91215  0.01724
## loudness         -0.39054     -0.04499  0.38655  7.26172  19.72907  0.17142
## tempo            -1.34280     -1.37723  1.91215 19.72907 985.98288 -0.17311
## valence          -0.00894      0.00850  0.01724  0.17142  -0.17311  0.03820
```

```
#Hotellings  - Independent Samples



HotellingsT2(Owned[,2:7], Unowned[,2:7])
```

```
##
##  Hotelling's two sample T2-test
##
## data:  Owned[, 2:7] and Unowned[, 2:7]
## T.2 = 5.8073, df1 = 6, df2 = 299, p-value = 9.697e-06
## alternative hypothesis: true location difference is not equal to c(0,0,0,0,0,0)
```

Hotellings indicates that these two groups have at least one set of unequal means among the predictors.

Train/Test Sets by Probabiltiy

```
set.seed(5)
n <- length(SelectData$Taylor_Owned)
train <- sample(n,round(n*0.8,0))
test <- setdiff(c(1:n),train)
```

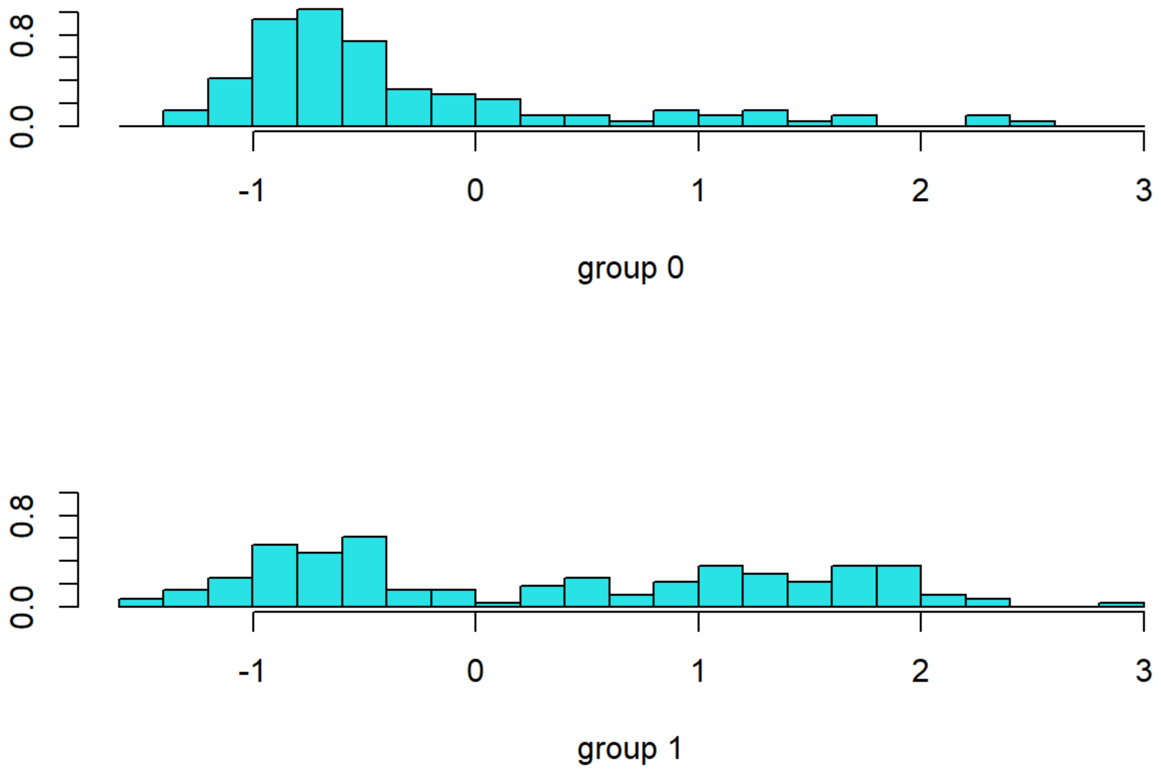
Parametric Plots (Unstable Error Rates)

LDA

```
set.seed(10)
lda.fit <- lda(Taylor_Owned~.-Index, data = SelectData, subset = train)
lda.fit
```

```
## Call:
## lda(Taylor_Owned ~ . - Index, data = SelectData, subset = train)
##
## Prior probabilities of groups:
##         0         1
## 0.4367347 0.5632653
##
## Group means:
##   acousticness danceability    energy  loudness    tempo   valence
## 0    0.2216257    0.5772523 0.6179533 -6.428991 123.8065 0.4165813
## 1    0.4118366    0.5645362 0.5325507 -8.049891 121.6551 0.3953500
##
## Coefficients of linear discriminants:
##                        LD1
## acousticness  1.897728e+00
## danceability -1.828043e+00
## energy        7.267158e-01
## loudness     -2.045308e-01
## tempo        -9.918248e-05
## valence       8.101370e-01
```

```
plot(lda.fit)
```



```
lda.pred <- predict(lda.fit, SelectData[test,])
lda.class <- lda.pred$class
table(lda.class, SelectData$Taylor_Owned[test])
```

```
##
## lda.class  0  1
##          0 20 11
##          1  9 21
```

```
mean(lda.class != SelectData$Taylor_Owned[test])
```

```
## [1] 0.3278689
```

Error for LDA 32.78 % (UNSTABLE in cross valdiation)

QDA

```
qda.fit <- qda(Taylor_Owned~ .-Index, data = SelectData, subset = train)
qda.fit
```

```
## Call:
## qda(Taylor_Owned ~ . - Index, data = SelectData, subset = train)
##
## Prior probabilities of groups:
##         0         1
## 0.4367347 0.5632653
##
## Group means:
##   acousticness danceability    energy  loudness    tempo   valence
## 0    0.2216257    0.5772523 0.6179533 -6.428991 123.8065 0.4165813
## 1    0.4118366    0.5645362 0.5325507 -8.049891 121.6551 0.3953500
```

```
qda.pred <- predict(qda.fit, SelectData[test,])
qda.class <- qda.pred$class
```

```
mean(qda.class != SelectData$Taylor_Owned[test])
```

```
## [1] 0.3770492
```

Error for QDA 37.7% This is not an improvement from QDA. (UNSTABLE in cross valdiation)

LOGISTIC

```
glm.auto <- glm(Taylor_Owned ~ .-Index, data = SelectData, subset = train, family = binomial)

summary(glm.auto)
```

```
## 
## Call:
## glm(formula = Taylor_Owned ~ . - Index, family = binomial, data = SelectData, 
##     subset = train)
## 
## Coefficients:
##                Estimate Std. Error z value Pr(>|z|)  
## (Intercept)  -9.110e-01  1.530e+00  -0.595   0.5516  
## acousticness  1.181e+00  6.645e-01   1.777   0.0756 .
## danceability -1.216e+00  1.453e+00  -0.837   0.4026  
## energy        4.766e-01  1.365e+00   0.349   0.7269  
## loudness     -1.392e-01  8.260e-02  -1.686   0.0919 .
## tempo         9.621e-07  4.644e-03   0.000   0.9998  
## valence       5.445e-01  9.463e-01   0.575   0.5650  
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## (Dispersion parameter for binomial family taken to be 1)
## 
##     Null deviance: 335.71  on 244  degrees of freedom
## Residual deviance: 312.01  on 238  degrees of freedom
## AIC: 326.01
## 
## Number of Fisher Scoring iterations: 4
```

```
glm.probs <- predict(glm.auto, SelectData[test,], type = "response")
head(glm.probs)
```

```
##        10        11        14        19        20        22 
## 0.4217274 0.4210208 0.3916391 0.4515779 0.5979546 0.5327819
```

```
glm.pred <- rep("0", nrow(SelectData[test,]))
glm.pred[glm.probs > 0.5] <- "1"

table(glm.pred, SelectData$Taylor_Owned[test])
```

```
##         
## glm.pred  0  1
##        0 20 11
##        1  9 21
```

```
lr.test.error <- mean(glm.pred != SelectData$Taylor_Owned[test])
lr.test.error
```

```
## [1] 0.3278689
```

PARAMETRIC Logistic Regression Test Error

Non Parametric Methods

KNN

```
SelectData.2 <- SelectData[,-1]
train.X <- SelectData.2[train,]
test.X <- SelectData.2[test,]

set.seed(20)

knn.pred2 <- knn(train.X, test.X, SelectData$Taylor_Owned[train], k = 2)
table(knn.pred2, SelectData$Taylor_Owned[test])
```

```
##          
## knn.pred2  0  1
##         0 15 10
##         1 14 22
```

```
mean(knn.pred2 != SelectData$Taylor_Owned[test])
```

```
## [1] 0.3934426
```

```
knn.pred4 <- knn(train.X, test.X, SelectData$Taylor_Owned[train], k = 4)
table(knn.pred4, SelectData$Taylor_Owned[test])
```

```
##          
## knn.pred4  0  1
##         0 16 11
##         1 13 21
```

```
mean(knn.pred4 != SelectData$Taylor_Owned[test])
```

```
## [1] 0.3934426
```

```
knn.pred6 <- knn(train.X, test.X, SelectData$Taylor_Owned[train], k = 6)
table(knn.pred6, SelectData$Taylor_Owned[test])
```

```
##          
## knn.pred6  0  1
##         0 21 12
##         1  8 20
```

```
mean(knn.pred6 != SelectData$Taylor_Owned[test])
```

```
## [1] 0.3278689
```

```
knn.pred8 <- knn(train.X, test.X, SelectData$Taylor_Owned[train], k = 8)
table(knn.pred8, SelectData$Taylor_Owned[test])
```

```
##
## knn.pred8  0  1
##         0 18 15
##         1 11 17
```

```
mean(knn.pred8 != SelectData$Taylor_Owned[test])
```

```
## [1] 0.4262295
```

```
knn.pred10 <- knn(train.X, test.X, SelectData$Taylor_Owned[train], k = 10)
table(knn.pred10, SelectData$Taylor_Owned[test])
```

```
##
## knn.pred10  0  1
##          0 16  6
##          1 13 26
```

```
mean(knn.pred10 != SelectData$Taylor_Owned[test])
```

```
## [1] 0.3114754
```

Best KNN parameter K=6, (1 overfits, 10 underfits)

CROSS VALIDATION with best KNN Parameter

```
set.seed(34)
n <- length(SelectData$Taylor_Owned)
knn.cv.error.5 <- rep(0,5)
n.test <- round(length(SelectData$Taylor_Owned)/5)

for (i in 1:5){
  # ordered test sequence
  cvtest <- seq((i-1)*n.test+1,min(i*n.test,n))
  #ordered train sequence
  cvtrain <- setdiff(c(1:n),cvtest)

  #KNN
  SelectData.2 <- SelectData[,-1]
  train.X <- SelectData.2[train,]
  test.X <- SelectData.2[test,]
  knn.pred <- knn(train.X, test.X, SelectData$Taylor_Owned[train], k = 6)
  knn.cv.error.5[i]<- mean(knn.pred!=SelectData$Taylor_Owned[test])
}

knn.cv.error.5
```

```
## [1] 0.3770492 0.3114754 0.3114754 0.2950820 0.3278689
```

```
mean(knn.cv.error.5)
```

```
## [1] 0.3245902
```

KNN Error Rate with 5-Fold Cross Validation 32.45%

Classifcation Trees

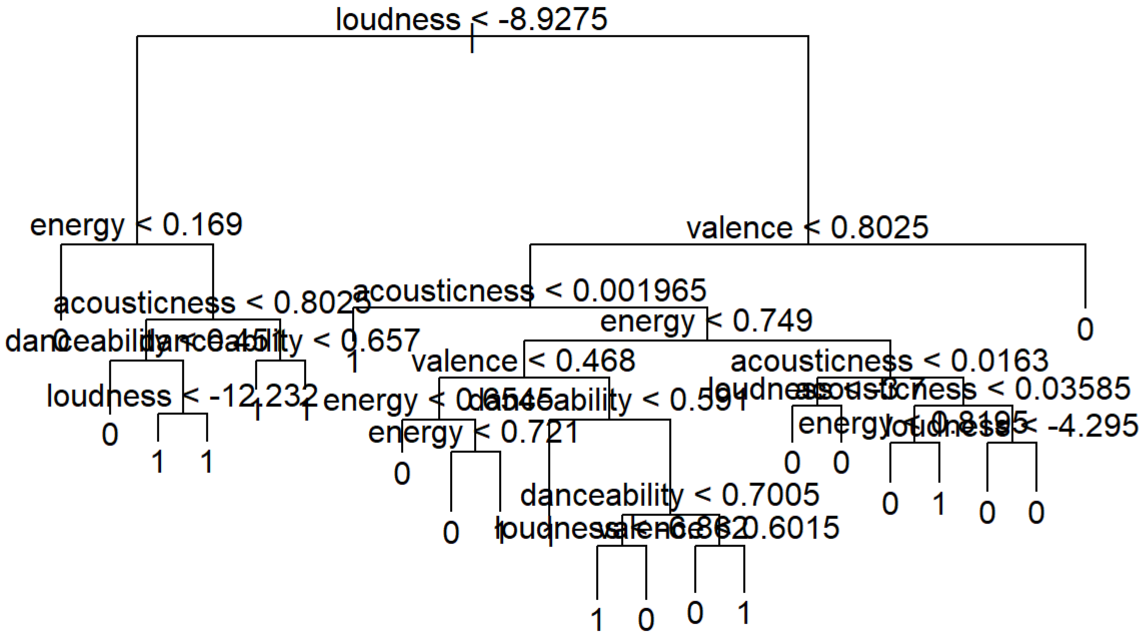
```
tree.taylor <- tree(Taylor_Owned ~ .-Index,SelectData[train,])

#Summary
summary(tree.taylor)
```

```
##
## Classification tree:
## tree(formula = Taylor_Owned ~ . - Index, data = SelectData[train,
##     ])
## Variables actually used in tree construction:
## [1] "loudness"     "energy"       "acousticness" "danceability" "valence"
## Number of terminal nodes:  22
## Residual mean deviance:  0.8071 = 180 / 223
## Misclassification error rate: 0.2286 = 56 / 245
```

22 NODES - TRAINING MISCLASSIFICATION RATE: 22.86%

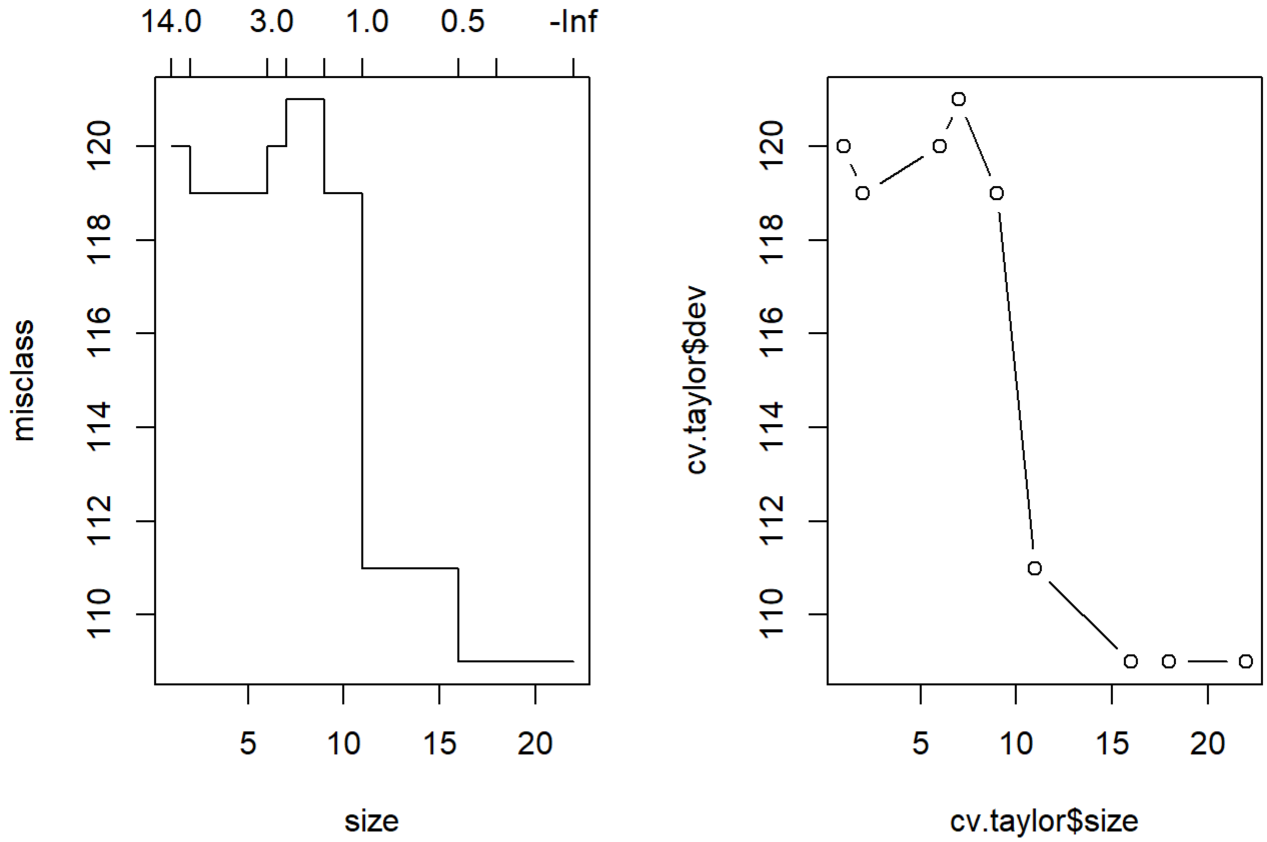
```
#Plot
plot(tree.taylor)
text(tree.taylor,pretty=0)
```



NODES: LOUDNESS –> Energy, Valence –> 0

```
# pruning
set.seed(121)
cv.taylor <- cv.tree(tree.taylor, FUN = prune.misclass)

par(mfrow=c(1,2))
plot(cv.taylor)
plot(cv.taylor$size,cv.taylor$dev,type="b")
```



BEST = 16

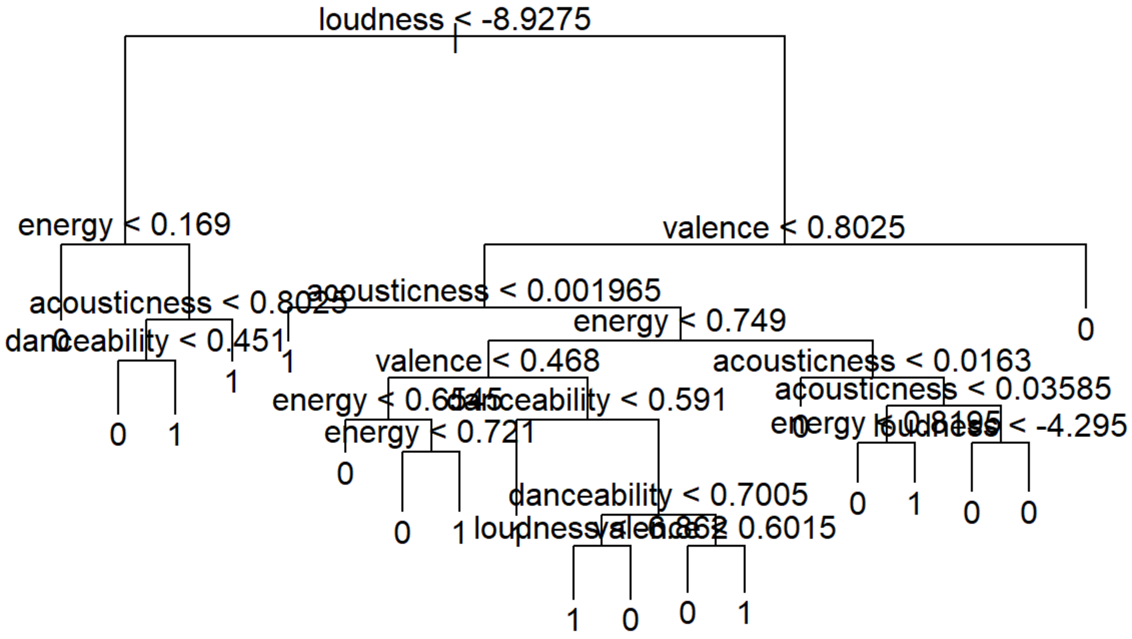
```
prune.taylor <- prune.tree(tree.taylor,best=16)

summary(prune.taylor)
```

```
##
## Classification tree:
## snip.tree(tree = tree.taylor, nodes = c(21L, 11L, 54L))
## Variables actually used in tree construction:
## [1] "loudness"     "energy"       "acousticness" "danceability" "valence"
## Number of terminal nodes:  19
## Residual mean deviance:  0.8529 = 192.8 / 226
## Misclassification error rate: 0.2286 = 56 / 245
```

19 NODES - TRAINING MISCLASSIFICATION RATE: 22.86%

```
plot(prune.taylor)
text(prune.taylor,pretty=0)
```



Not much pruning, NODES: LOUDNESS –> Energy, Valence –> 0

```
tree.pred <- predict(prune.taylor,SelectData[test,],type="class")
table(tree.pred,SelectData$Taylor_Owned[test])
```

```
##
## tree.pred  0  1
##          0 23 18
##          1  6 14
```

```
#Error
mean (tree.pred!= SelectData$Taylor_Owned[test])
```

```
## [1] 0.3934426
```

TEST ERROR RATE: 39.34%

Random Forest

```
set.seed(131)
rf.taylor <- randomForest(Taylor_Owned ~ .-Index,data= SelectData, subset= train, mtry= 6, ntree
= 5000)
rf.pred <- predict(rf.taylor,SelectData[test,],type="class")
table(rf.pred,SelectData$Taylor_Owned[test])
```

```
##
## rf.pred  0  1
##       0 19 13
##       1 10 19
```

```
#Error
mean (rf.pred!= SelectData$Taylor_Owned[test])
```

```
## [1] 0.3770492
```

TEST ERROR RATE: 37.7%