



Data Visualization & Mining

Data Visualization and Mining on Music Track Features

Authors' Name	Student ID
Islam Anwar	201900337

FACULTY OF SCIENCE & INNOVATION,
THE UNIVERSITIES OF CANADA

December 25, 2022

Abstract

Data visualization and mining are becoming integral aspects of business and data science. Extraction of data and insights has never been more important to guide more effective business decisions and learning more about the data. Various techniques are explored in this paper including principal component analysis, univariate, bivariate, and multivariate outlier detection methods, cluster analysis, and more. Results indicate that there is a significant amount of data and information to be extracted from the data. However, there is still untapped infinite potential regarding using other information-extracting techniques. This paper introduces the basic techniques for data mining while also paving the way for improved future work.

Principal Component Analysis — Multivariate Outlier Detection — Cluster Analysis

Contents

1	Introduction	1
2	Literature Review	1
3	Data Source	2
3.1	Types of Observations	2
3.2	Variable Types and Units of Measurement	2
4	Data Visualization & Mining	3
4.1	Single-Variable Visualization	3
4.2	Linear Discriminant Analysis	4
4.3	Cluster Analysis	4
4.3.1	Hierarchical Clustering	4
4.3.2	K-Means Clustering	5
4.4	Hotelling's T^2 Test	6
4.5	Principal Component Analysis	7
4.6	Multidimensional Scaling	8
4.7	Canonical Correlation Analysis	8
4.8	Measures of Association	9
4.9	Outlier Detection	9
4.9.1	Univariate Analysis	9
4.9.2	Bivariate Analysis	9
4.9.3	Multivariate Analysis	10
4.9.4	Revisiting previous techniques post-outlier removal	11
5	Conclusions & Discussion	13
5.1	Discussion & Future Work	13
6	Appendix	13
6.1	Annotated Computer Output (Tables & Figures)	13

1 Introduction

Data visualization and mining are one of the cornerstones of data science [1]. The emergence of data-driven analytics and big data processing has introduced a revolutionary industry. An industry of data processing methodologies, extraction of insights, analysis of noise, and many other applications.

Businesses worldwide aim to use data analysis to extract important business metrics and Key Performance Indicators (KPIs) that may not be visible to the human eye simply by observing the data. But by diving deeper into the nature of that data, visualizing relationships among features, and most importantly, explaining why these relationships are present.

Some questions that can be answered using data visualization & mining:

- What are the relationships between the variables?
- Is there any underlying information that can be extracted using visualization techniques?
- Are there some variables that are more significant than others?

This paper aims to produce deep and clear insights into a large data set by extracting the most influential features. Then performing data preprocessing and visualization. The structure of this paper is as follows: Section 2 highlights essential data mining techniques and some literature reviews. Section 3 discusses the data set of choice in detail. This includes the type of observations and variable types. Section 4 is primarily focused on the data visualization and mining portion of the paper and illustrates the preprocessing techniques performed on the raw data. Data mining techniques are discussed in detail including cluster analysis, linear discriminant analysis, and hypothesis testing. Conclusions and discussions are showcased in Section 5, regarding the main insights that were extracted and any suggestions for future work. Section 6 is an appendix of the code used in this paper.

2 Literature Review

There is a multitude of data mining techniques that have yielded exceptional results over the past few years of research. Among them is cluster analysis, which is used to group observations according to similar or dissimilar characteristics in higher dimensions. This can be a valuable tool to not only plot data in hyperspaces, but also to identify some of the most influential features amongst our data set. Another technique is association rules, which work to discover relationships or correlations between a set of observations within a data set to create deeper insights. Associations can be constructed as single-dimensional or multi-dimensional associations based on the nature of the features and their correlation values.

Furthermore, data cleaning and visualizations are among the most valuable techniques for any data scientist. Tasks such as checking for null values, converting data types,

removing irrelevant data, and computing the most influential features. These are a few of the most crucial tasks of data preprocessing before building visuals and representing our data. Data visualization can come in many forms from comparison charts, heat maps to contour plots and parallel coordinates plots. The ability to discern which visualizations are appropriate based on the data type of each feature is an important skill.

Cluster analysis is one of the most important data mining techniques for identifying underlying patterns among data. Various steps are taken in cluster analysis to produce effective insights and similar groups. The first step in the process is pattern representation, which involves feature extraction or feature selection techniques to obtain the most valuable features that differentiate between observations and group observations of similar characteristics. The next step involves assigning a similarity measure to quantify how close two data points are. Once a similarity measure is assigned, a clustering algorithm such as k-means or hierarchical clustering is deployed to group the data points accordingly. Followed by certain cluster validity measures and visualization to observe the accuracy of the clustering algorithm [2].

3 Data Source

The chosen data set for this report is the Spotify 1.2M+ Songs obtained from the Spotify API. This data set features over 1.2 million songs, each with their own unique features and values. However, due to the massive computational requirement of analysing 1.2 million observations at 24 variables, we will use a sample of 5000 observations. The reasoning behind choosing this data set is to understand the key features that differentiate different types of music, ranging from genre, danceability and other factors. The purpose of this experiment is to perform data visualization and mining techniques on our data set to yield deeper insights.

3.1 Types of Observations

As illustrated in Fig. (17), there are 24 features within this data source with varying types including liveness, speechiness, release_date, etc. Observations of this data set consist of unique identifiers for each row, with a song title. Each row is also accompanied by several other details including album_id, track_number, explicit, valence, tempo, time_signature, and more.

3.2 Variable Types and Units of Measurement

Fig. (1) displays the data type of each respective feature of our data set. Some features such as the ids will not be very beneficial to our work. However, other features such as the name of an album, artist names, and musical characteristics that are mostly integer data types will be very useful to perform filtering operations and cluster analysis. Regarding having a response variable, there is no set response variable, but depending

on the visualization technique or mining method performed. We may assign one of the features to be the response variable to conduct certain statistical techniques such as Hierarchical Clustering (HC) or Fisher Linear Discriminant Analysis (FLDA).

```
(r)
library("tidyverse")
str(musicData)

'data.frame': 1204025 obs. of 24 variables:
 $ id      : chr "7lmeHLH8e4nmzxuc0HDjk" "1wsRitFRrtwyEap10q22o8" "1hr0fIFK2qRG3f3RF70pb7" "2lbaSgTS0d07MTULAXlTw0" ...
 $ name    : chr "Testify" "Guerrilla Radio" "Calm Like a Bomb" "Mic Check" ...
 $ album   : chr "The Battle Of Los Angeles" "The Battle Of Los Angeles" "The Battle Of Los Angeles" "The Battle Of Los Angeles" ...
 $ album_id : chr "2eia0mywFgoHuttJytcxgx" "2eia0mywFgoHuttJytcxgx" "2eia0mywFgoHuttJytcxgx" "2eia0mywFgoHuttJytcxgx" ...
 $ artists : chr "['Rage Against The Machine']" "['Rage Against The Machine']" "['Rage Against The Machine']" "['Rage Against The Machine']" ...
 $ artist_ids : chr "['2d0hyoQ5ynBnkVAbjKORj']" "['2d0hyoQ5ynBnkVAbjKORj']" "['2d0hyoQ5ynBnkVAbjKORj']" "['2d0hyoQ5ynBnkVAbjKORj']" ...
 $ track_number : int 1 2 3 4 5 6 7 8 9 10 ...
 $ disc_number : int 1 1 1 1 1 1 1 1 1 1 ...
 $ explicit  : chr "False" "True" "False" "True" ...
 $ danceability : num 0.47 0.599 0.315 0.44 0.426 0.298 0.417 0.277 0.441 0.448 ...
 $ energy     : num 0.978 0.957 0.97 0.967 0.929 0.848 0.976 0.873 0.882 0.861 ...
 $ key        : int 7 11 7 11 2 2 9 11 7 9 ...
 $ loudness   : num -5.4 -5.76 -5.42 -5.83 -6.73 ...
 $ mode       : int 1 1 1 0 1 1 1 0 1 1 ...
 $ speechiness : num 0.0727 0.188 0.483 0.237 0.0701 0.0727 0.175 0.0883 0.044 0.0676 ...
 $ acousticness : num 0.0261 0.0129 0.0234 0.163 0.00162 0.0538 0.000427 0.00694 0.0195 0.00306 ...
 $ instrumentalness : num 1.09e-05 7.06e-05 2.03e-06 3.64e-06 1.05e-01 1.52e-03 1.34e-04 5.40e-05 6.84e-03 0.00 ...
 $ liveness    : num 0.356 0.155 0.122 0.121 0.0789 0.201 0.107 0.188 0.15 0.0987 ...
 $ valence     : num 0.503 0.489 0.37 0.574 0.539 0.194 0.483 0.618 0.418 0.761 ...
 $ tempo       : num 117.9 103.7 149.7 96.8 127.1 ...
 $ duration_ms : int 210133 206200 298893 213640 205600 280960 202040 228093 151573 224933 ...
 $ time_signature : num 4 4 4 4 4 4 4 4 4 ...
 $ year        : int 1999 1999 1999 1999 1999 1999 1999 1999 1999 1999 ...
 $ release_date : chr "1999-11-02" "1999-11-02" "1999-11-02" "1999-11-02" ...
```

Figure 1: Variable Data Types

4 Data Visualization & Mining

4.1 Single-Variable Visualization

To perform analysis on our data set, we must understand the nature of the observations at hand and how we can use that to our advantage. This is where data visualization thrives. It is an essential technique that helps us understand our data and perform more meaningful analyses. This section covers a single variable analysis of all numerical features including tempo, key, acousticness, and others.

As shown in Fig. (20) to Fig. (39), there is a large variance in each of our variables. Box plots were created to identify outliers and observe the spread of the data. Some variables such as acousticness and instrumentalness had a very large spread in their values while others like tempo and speechiness had very little spread with an assortment of outliers present.

Regarding distributions of our variables, Fig. (20) to Fig. (39) also displayed the individual histograms of each of our numerical variables. A few features such as tempo and danceability had normally distributed data. On the other hand, some variables were left and right skewed in their distribution such as loudness and liveness respectively. Other features showed random distributions, which could be due to the massive variety of observations in our music data set.

Concerning correlation, Fig. (40) displays the correlation matrix of our numerical variables. Among some of the most notable correlations are between energy and loudness with a strong positive correlation of 0.82. Conversely, there is a strong negative correlation between energy and acousticness of -0.82. These results are expected as loud music tends to instill energy and lacks acousticness.

4.2 Linear Discriminant Analysis

To further explore our dataset, we performed Linear Discriminant Analysis (LDA) to be able to predict the key of a new observation based on the current data that we possess. We began this process by first removing all numerical variables that did not provide any information or value to our analysis including the tracking number of a song within an album, the release date, and the year. Thus, producing a dataset of 10 features and our "key" label for a total of 11 columns as shown in Fig. (41).

The next step to performing LDA is to scale the features. Fig. (42) and (43) showcase the scaling procedure and how we verified that the standard deviation of all our features was set to 1. Furthermore, we split our dataset using a train-test split approach with a distribution of 70% and 30% respectively. Furthermore, we created our LDA model and fit the model onto our training data with the key as our label. Fig. (44) shows our model summary including the prior probabilities, group means of each respective feature, as well as the proportion of the trace for each linear discriminant.

We then attempted to test our model by predicting the label of some of our observations. As shown in Fig. (45), we predicted the first 6 observations of our data, and Fig. (46) evaluated the posterior probabilities of the first 6 observations belonging to each potential key (with a total of 11 possible keys). However, upon evaluating the true accuracy of our LDA model, we obtained an average accuracy of only 0.15 as illustrated in Fig. (47), which indicates that this model is very poor at predicting the key of a song.

4.3 Cluster Analysis

4.3.1 Hierarchical Clustering

The first clustering method we are using is Hierarchical Clustering (HC) and in particular agglomerative HC. Agglomerative HC is a clustering technique that assigns each observation a cluster of its own. Then using a chosen distance measure, it assigns each data point to the closest cluster and repeats the process until all points have been assigned to a single cluster. However, realistically a single cluster would not be useful to represent our observations therefore, we use a diagram known as a dendrogram to determine an appropriate number of clusters for our data.

Our first action is to remove the label from the dataset as shown in Fig. () to cluster according to the features only. Our chosen distance measure between observations was the euclidean distance as shown in Equation (1).

$$d(p, q) = \sqrt{\sum_{i=1}^n (q_i - p_i)^2} \quad (1)$$

Fig. (49) showcases our HC model with a complete distance measure between clusters. This method was chosen because it tends to provide compact clusters with a small variance within the cluster. Furthermore, we wanted to obtain the optimal number of

clusters for our data. Fig. (2) is an illustration of our dendrogram and base of the distribution, it would seem that our appropriate number of clusters should range from 3 to 5 due to the high variability at the top of the dendrogram between clusters.

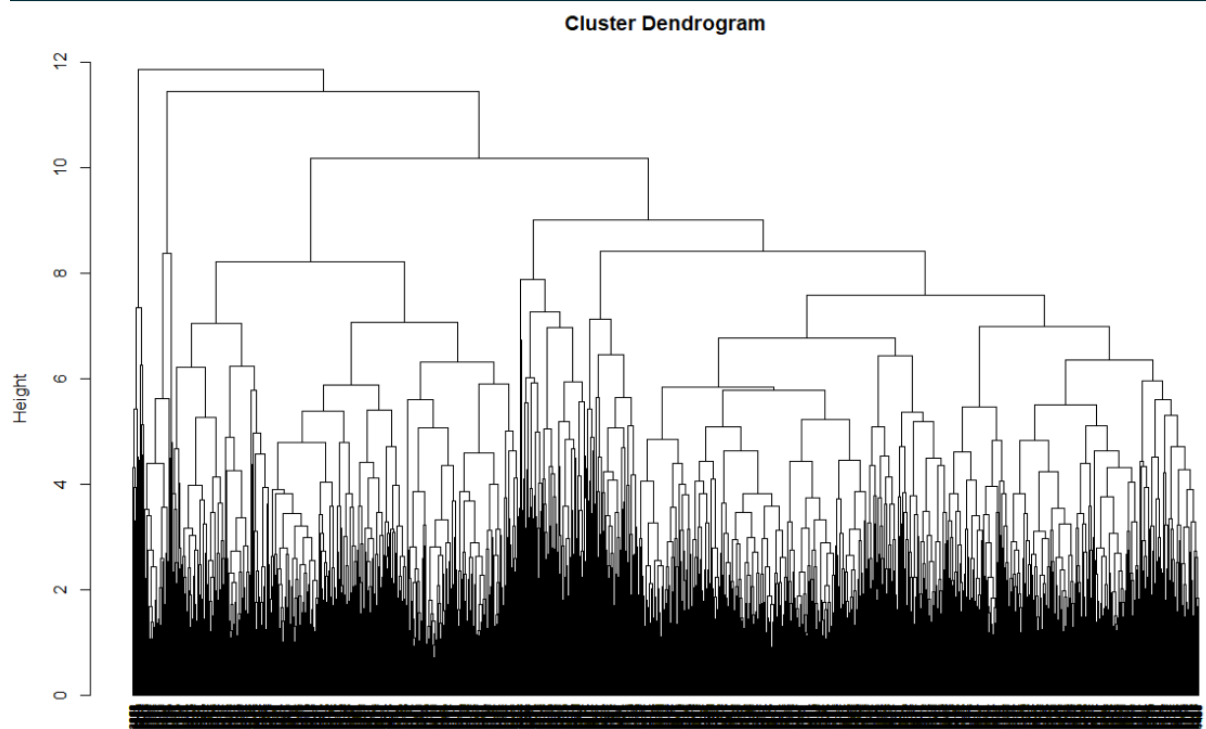


Figure 2: Dendrogram for HC

4.3.2 K-Means Clustering

Our next clustering method is k-means clustering. K-means clustering works with the assumption that the number of clusters is known before clustering. This can be identified by running k-means with a different number of clusters and constructing a plot where the elbow method can be applied. Fig. (50) shows the plot using different numbers of clusters and applying the elbow method, it would be feasible to assume that $k=5$.

Once k-means begins, it randomly initializes cluster centers for each cluster and then assigns each observation to the closest corresponding cluster center. This process is repeated and cluster centers are recalculated until convergence occurs, or a criterion is met that stops the clustering. From our results, k-means iterated 6 times as shown in Fig. (51). Regarding the evaluation of our k-means clustering, we obtained a goodness of fit of approximately 43%, which is considered a substandard value for the accuracy of clusters.

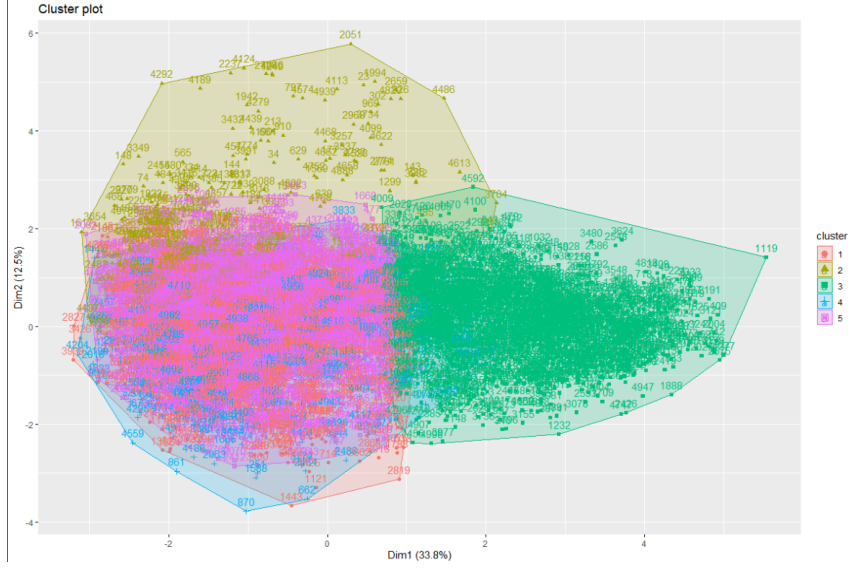


Figure 3: K-Means Cluster Output

Fig. (3) shows our cluster output, and it is clear that there is a significant number of observations that belong to more than one cluster. Furthermore, the shapes of our clusters are not uniform, which can lead us to conclude that perhaps HC and k-means clustering were not the most effective clustering techniques as the nature of our data is density-based.

4.4 Hotelling's T^2 Test

Our next area of analysis is Hotelling's T^2 test that calculates the elliptical distance between the two sample means as shown in Equation (2). The benefit of using hotelling's test is that we can relate it to the F test commonly used in statistics as shown in Equation (3).

We first performed a normality check using the Shapiro test on our hypothesis data to ensure normality as well as running a determinant check on our data. We computed the determinant of the variance-covariance matrix of the hypothesis data to ensure our data was represented as a non-singular matrix. Both Fig. (41) and (42) illustrate our preprocessing steps for hotelling's two-sample mean test.

$$T^2 = (\bar{x} - \bar{y})^T \left[\frac{n+m}{nm} S_{pooled} \right] (\bar{x} - \bar{y}) \quad (2)$$

$$\frac{n+m-p-1}{p(n+m-2)} T^2 = F_{p, n+m-p-1} \quad (3)$$

From Fig. (4), we can see the result of performing the hotelling's test using the explicit feature as our label from our original dataset. Explicit only consists of two categorical values: TRUE and FALSE. Therefore, we performed hotelling's test on the equality of

means between when explicit was equal to true and false for all numerical features. From our results, we can see the different degrees of freedom as well as the p-value, which is less than the default $\alpha=0.05$. This indicates that the test is statistically significant and we reject the null hypothesis, and the difference between all the combined features is 0 between all of them.

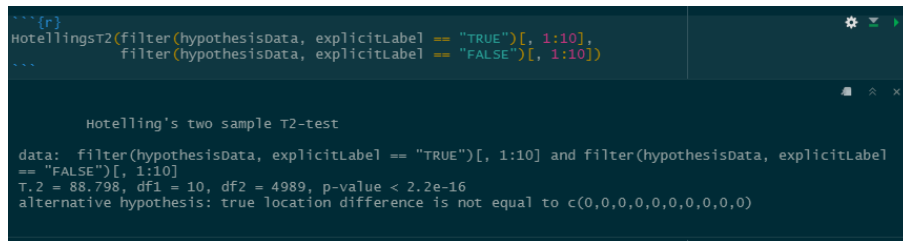


Figure 4: Hotelling's T^2 Test

4.5 Principal Component Analysis

Among one the most vital exploratory data analysis tools are Principal Component Analysis (PCA). It is used primarily to reduce the number of dimensions of a multivariate dataset. This is performed by first normalizing our data. Then we retrieve the principal components by computing the eigenvalues and eigenvectors of each numerical variable in our data.

Principal components were computed and Fig. (6) illustrates the proportion of variances among 11 different components. Furthermore, the scree plot in Fig. (5) surmises that dimensionality reduction through PCA is not feasible because the proportion of variances among the components is almost evenly distributed. Therefore, choosing a subset of components would significantly impact the retained information from our data.

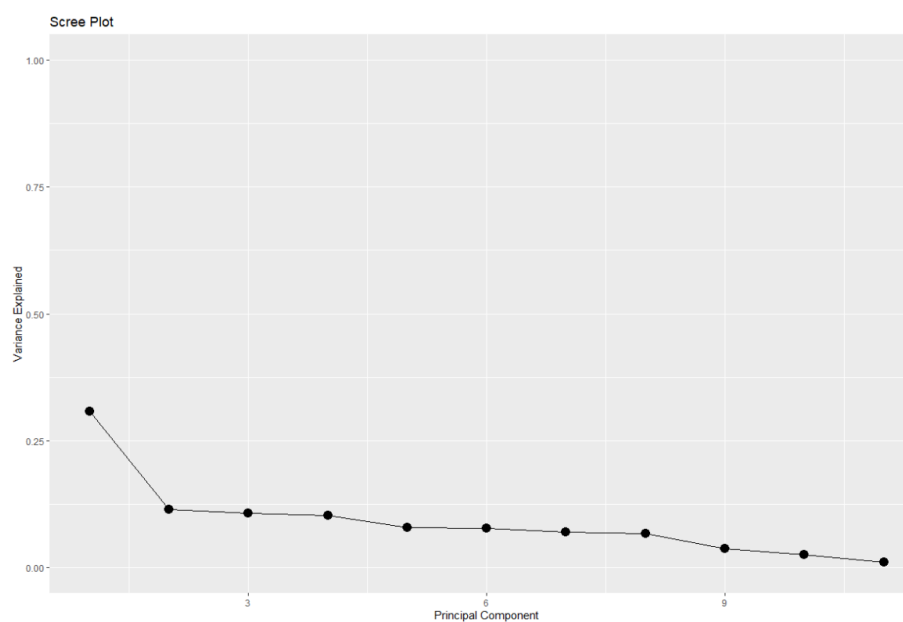


Figure 5: Scree Plot

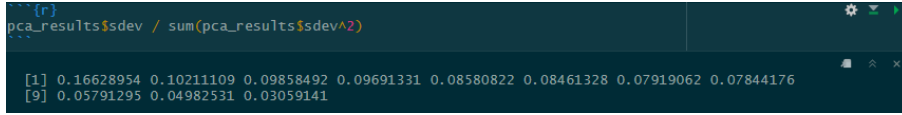


Figure 6: Principal Component Analysis Variance Proportions

4.6 Multidimensional Scaling

Another facet of multivariate analysis is Multi-Dimensional Scaling (MDS). MDS is used to measure similarity or dissimilarity between observations to be visualized in two dimensions. It is fed a distance matrix between observations and computes the corresponding eigenvalues and eigenvectors, also known as principal coordinates.

Our experiment displayed various outputs including the coordinates of the points, the computed eigenvalues, a centered distance matrix, and a numeric vector of length 2 that gives the decreasing order of eigenvalues as shown in Fig. (7).



Figure 7: MDS Goodness of Fit

4.7 Canonical Correlation Analysis

Canonical correlation is an analysis technique used to measure relationships and correlations between two sets of data. A 2x2 matrix is generated that contains the correlations of each dataset with itself in the main diagonal, and the minor diagonal contains the correlations between the datasets.

Our dataset is split evenly into 2500 observations to two datasets for this experiment by adjusting the danceability variable to 0.36. Fig. (8) and (9) display the XY correlation between our two datasets.

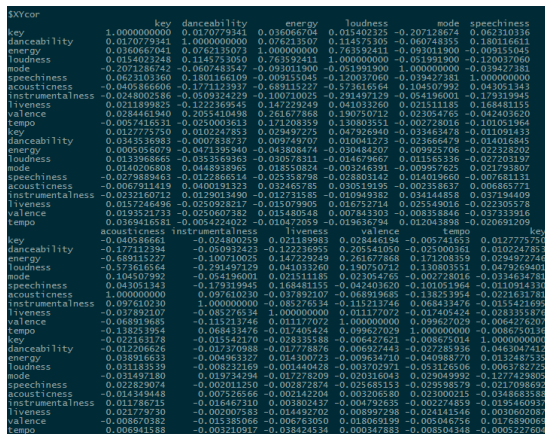


Figure 8: XY Correlation 1

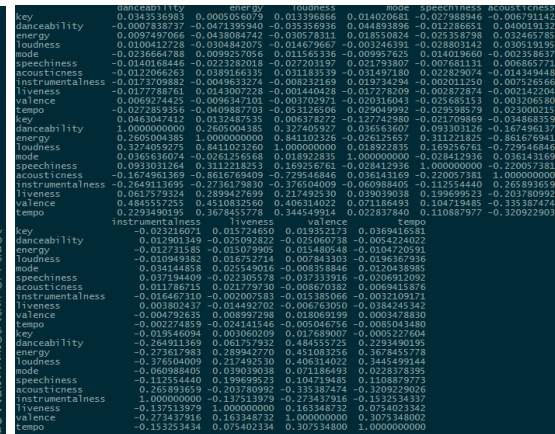


Figure 9: XY Correlation 2

4.8 Measures of Association

Due to our dataset containing very few categorical variables, we will attempt to find measures of association between variables using other techniques such as the chi-squared independence test. Based on the results of the chi-squared independence test, Fig. (10) illustrates the correlation between artists and whether their music is explicit, which shows independence between artists and if their music is usually explicit or not.

```
{r}
chisq.test(musicData$artists, musicData$explicit)

Warning in chisq.test(musicData$artists, musicData$explicit) :
  Chi-squared approximation may be incorrect

Pearson's Chi-squared test

data: musicData$artists and musicData$explicit
X-squared = 4948.4, df = 4573, p-value = 6.417e-05
```

Figure 10: Independence test of artist and explicit

4.9 Outlier Detection

This section highlights identifying outliers for univariate, bivariate, and multivariate scenarios. Furthermore, we perform the above sections again after outlier detection and measure changes in results.

4.9.1 Univariate Analysis

Regarding univariate analysis, the box-plots included in Fig. (20) to (39) show the outliers present in any of our numerical variables. There are a few outliers present in the tempo variable as well as liveness, loudness, and speechiness.

4.9.2 Bivariate Analysis

Concerning bivariate analysis, we plotted the scatter matrix between each pair of numerical variables. According to our results in Fig. (11), we can see that many of the pairs do not follow any linear representation in their scatter plots, while a few pairs such as the loudness and energy contain some linearity. In the grand scheme, the outliers are not easy to identify using these pairs matrix. A better way to identify outliers numerically is in our multivariate analysis, which will be discussed briefly.

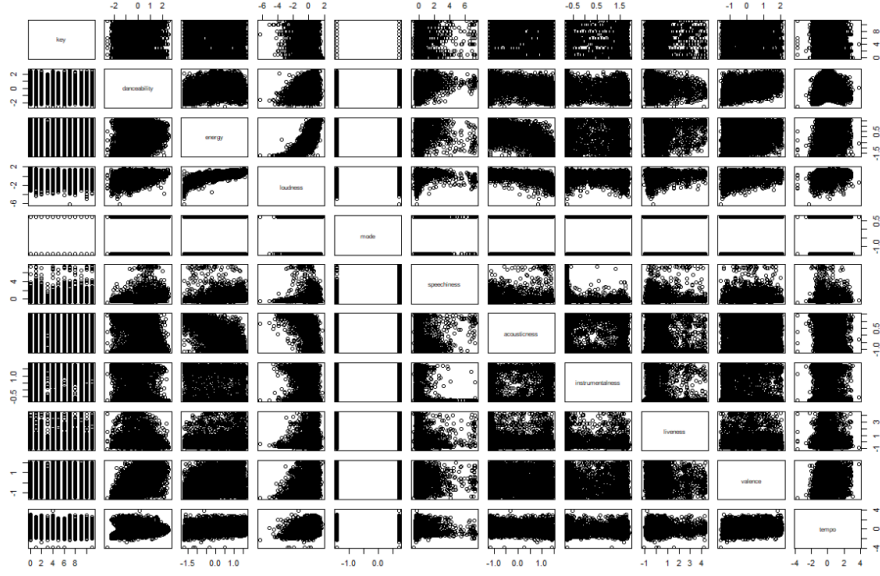


Figure 11: Scatter Matrix for Outlier Check

4.9.3 Multivariate Analysis

For multivariate analysis, we performed the Mahalanobis distance for each observation. Based on Fig. (12), we can interpret that approximately 635-650 outliers are present in our data. This may seem like a poor decision to remove so many observations from our data. However, it would be interesting to see the effect of removing these outliers (regardless of their influence) on the results of our previous techniques. Although they would be considered meaningful outliers because of the nature of our data having varying values in a multitude of musical features, we will remove these observations and carry out the previous techniques without the outliers to observe if there are any significant changes in the performance of our approaches. We remove observations with a p-value less than an $\alpha=0.05$ and then form a new dataset for repeating our techniques.

```

[1] music_num[outlier] == "no"
[1] music_num[outlier][music_numip < 0.05] <- "yes"

[1]
[1] music_num[outlier][music_numip < 0.05]

[1] "yes" "yes" "yes" "yes" "yes" "yes" "yes" "yes" "yes" "yes" "yes" "yes" "yes" "yes" "yes"
[17] "yes" "yes" "yes" "yes" "yes" "yes" "yes" "yes" "yes" "yes" "yes" "yes" "yes" "yes" "yes"
[33] "yes" "yes" "yes" "yes" "yes" "yes" "yes" "yes" "yes" "yes" "yes" "yes" "yes" "yes" "yes"
[49] "yes" "yes" "yes" "yes" "yes" "yes" "yes" "yes" "yes" "yes" "yes" "yes" "yes" "yes" "yes"
[65] "yes" "yes" "yes" "yes" "yes" "yes" "yes" "yes" "yes" "yes" "yes" "yes" "yes" "yes" "yes"
[81] "yes" "yes" "yes" "yes" "yes" "yes" "yes" "yes" "yes" "yes" "yes" "yes" "yes" "yes" "yes"
[97] "yes" "yes" "yes" "yes" "yes" "yes" "yes" "yes" "yes" "yes" "yes" "yes" "yes" "yes" "yes"
[113] "yes" "yes" "yes" "yes" "yes" "yes" "yes" "yes" "yes" "yes" "yes" "yes" "yes" "yes" "yes"
[129] "yes" "yes" "yes" "yes" "yes" "yes" "yes" "yes" "yes" "yes" "yes" "yes" "yes" "yes" "yes"
[145] "yes" "yes" "yes" "yes" "yes" "yes" "yes" "yes" "yes" "yes" "yes" "yes" "yes" "yes" "yes"
[161] "yes" "yes" "yes" "yes" "yes" "yes" "yes" "yes" "yes" "yes" "yes" "yes" "yes" "yes" "yes"
[177] "yes" "yes" "yes" "yes" "yes" "yes" "yes" "yes" "yes" "yes" "yes" "yes" "yes" "yes" "yes"
[193] "yes" "yes" "yes" "yes" "yes" "yes" "yes" "yes" "yes" "yes" "yes" "yes" "yes" "yes" "yes"
[209] "yes" "yes" "yes" "yes" "yes" "yes" "yes" "yes" "yes" "yes" "yes" "yes" "yes" "yes" "yes"
[225] "yes" "yes" "yes" "yes" "yes" "yes" "yes" "yes" "yes" "yes" "yes" "yes" "yes" "yes" "yes"
[241] "yes" "yes" "yes" "yes" "yes" "yes" "yes" "yes" "yes" "yes" "yes" "yes" "yes" "yes" "yes"
[257] "yes" "yes" "yes" "yes" "yes" "yes" "yes" "yes" "yes" "yes" "yes" "yes" "yes" "yes" "yes"
[273] "yes" "yes" "yes" "yes" "yes" "yes" "yes" "yes" "yes" "yes" "yes" "yes" "yes" "yes" "yes"
[289] "yes" "yes" "yes" "yes" "yes" "yes" "yes" "yes" "yes" "yes" "yes" "yes" "yes" "yes" "yes"
[305] "yes" "yes" "yes" "yes" "yes" "yes" "yes" "yes" "yes" "yes" "yes" "yes" "yes" "yes" "yes"
[321] "yes" "yes" "yes" "yes" "yes" "yes" "yes" "yes" "yes" "yes" "yes" "yes" "yes" "yes" "yes"
[337] "yes" "yes" "yes" "yes" "yes" "yes" "yes" "yes" "yes" "yes" "yes" "yes" "yes" "yes" "yes"
[353] "yes" "yes" "yes" "yes" "yes" "yes" "yes" "yes" "yes" "yes" "yes" "yes" "yes" "yes" "yes"
[369] "yes" "yes" "yes" "yes" "yes" "yes" "yes" "yes" "yes" "yes" "yes" "yes" "yes" "yes" "yes"
[385] "yes" "yes" "yes" "yes" "yes" "yes" "yes" "yes" "yes" "yes" "yes" "yes" "yes" "yes" "yes"
[401] "yes" "yes" "yes" "yes" "yes" "yes" "yes" "yes" "yes" "yes" "yes" "yes" "yes" "yes" "yes"
[417] "yes" "yes" "yes" "yes" "yes" "yes" "yes" "yes" "yes" "yes" "yes" "yes" "yes" "yes" "yes"
[433] "yes" "yes" "yes" "yes" "yes" "yes" "yes" "yes" "yes" "yes" "yes" "yes" "yes" "yes" "yes"
[449] "yes" "yes" "yes" "yes" "yes" "yes" "yes" "yes" "yes" "yes" "yes" "yes" "yes" "yes" "yes"
[465] "yes" "yes" "yes" "yes" "yes" "yes" "yes" "yes" "yes" "yes" "yes" "yes" "yes" "yes" "yes"
[481] "yes" "yes" "yes" "yes" "yes" "yes" "yes" "yes" "yes" "yes" "yes" "yes" "yes" "yes" "yes"
[497] "yes" "yes" "yes" "yes" "yes" "yes" "yes" "yes" "yes" "yes" "yes" "yes" "yes" "yes" "yes"
[513] "yes" "yes" "yes" "yes" "yes" "yes" "yes" "yes" "yes" "yes" "yes" "yes" "yes" "yes" "yes"
[529] "yes" "yes" "yes" "yes" "yes" "yes" "yes" "yes" "yes" "yes" "yes" "yes" "yes" "yes" "yes"
[545] "yes" "yes" "yes" "yes" "yes" "yes" "yes" "yes" "yes" "yes" "yes" "yes" "yes" "yes" "yes"
[561] "yes" "yes" "yes" "yes" "yes" "yes" "yes" "yes" "yes" "yes" "yes" "yes" "yes" "yes" "yes"
[577] "yes" "yes" "yes" "yes" "yes" "yes" "yes" "yes" "yes" "yes" "yes" "yes" "yes" "yes" "yes"
[593] "yes" "yes" "yes" "yes" "yes" "yes" "yes" "yes" "yes" "yes" "yes" "yes" "yes" "yes" "yes"
[609] "yes" "yes" "yes" "yes" "yes" "yes" "yes" "yes" "yes" "yes" "yes" "yes" "yes" "yes" "yes"
[625] "yes" "yes" "yes" "yes" "yes" "yes" "yes" "yes" "yes" "yes" "yes" "yes" "yes" "yes" "yes"

```

Figure 12: Outlier detection using mahalanobis distance

4.9.4 Revisiting previous techniques post-outlier removal

Regarding LDA, after removing the outliers we had 4,360 observations and performed the aforementioned experiment again. Results in Fig. (54) and (55) show that the information among the linear discriminants was more spread out towards some of the first 2 or 3 linear discriminants as opposed to our previous model which had the information almost evenly distributed among the discriminants.

Concerning Hierarchical clustering, results indicated in the following dendrogram show that there is more variability among the clusters as we get closer to combining them into a single cluster. This is an improvement on the previous experiment as we only had 4 to 5 clusters, but this dendrogram is more promising as we may choose approximately 5-7 clusters as shown in Fig. (13).

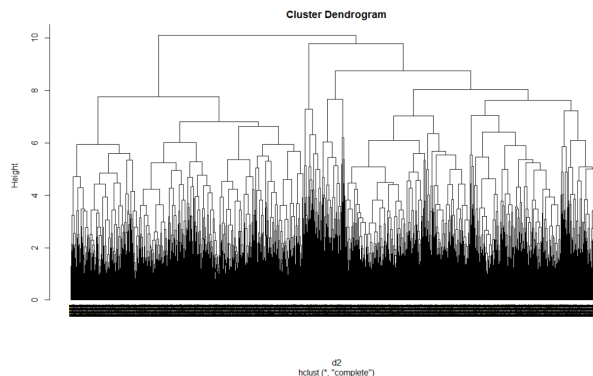


Figure 13: Hierarchical clustering dendrogram

For k-means clustering, we found that the clustering was best performed at k=6 due to the elbow method in Fig. (56). The goodness of fit of this model was 49%, which is an improvement to the 43% of the previous model. This could be due to the removal of

outliers being impactful to k-means clustering as it relies heavily on the use of central measures such as the mean. Fig. (14) illustrates the output of the new model.

```
within cluster sum of squares by cluster:
[1] 4741.703 7005.090 5026.952 5019.826 2096.135
(between_SS / total_SS = 45.2 %)
```

Figure 14: K-means clustering output

The hotelling test was performed again using the categorical variable 'Explicit' of values TRUE and FALSE. Results illustrated that with a significance level of $\alpha=0.05$, the test was statistically significant and rejected the null hypothesis. Outliers had no significant effect on the hotelling test.

Regarding PCA, Fig. (15) shows the graph of the cumulative variance of each respective principal component. As the graph shows, the result of removing outliers did not affect principal components, the cumulative variance was very evenly distributed among the components as in our first experiment.

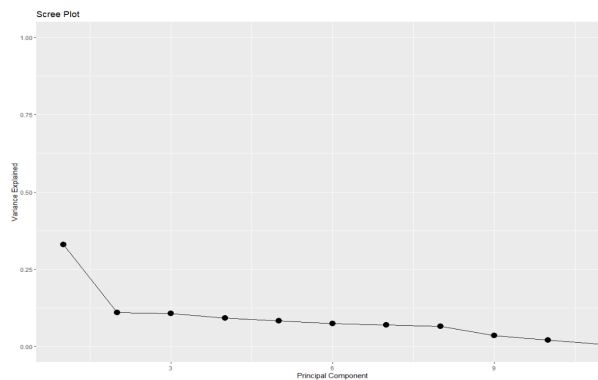


Figure 15: PCA Scree plot after outlier removal

For MDS, we experimented using the new distance matrix for our observations excluding outliers, and ended up with a goodness of fit of approximately 47.9%, a very small marginal improvement on the original trial. Fig. (57) displays the goodness of fit for our second attempt.

For canonical correlation, Fig. (16) shows a maximum correlation of 0.13 between the two datasets split, this is a small improvement yet not too significant as to suggest that our outliers were a strong impact on our original experiments.

```
(r)
cc2 <- matcor(dataset1_cancor2, dataset2_cancor2)
cc2 <- cc(dataset1_cancor2, dataset2_cancor2)
cc2$cor
[1] 0.130683642 0.108805514 0.100459524 0.087234862 0.075781104 0.062711651 0.055348214 0.035329161
[9] 0.024523821 0.013834016 0.006196697
```

Figure 16: Cancor after outlier removal

5 Conclusions & Discussion

5.1 Discussion & Future Work

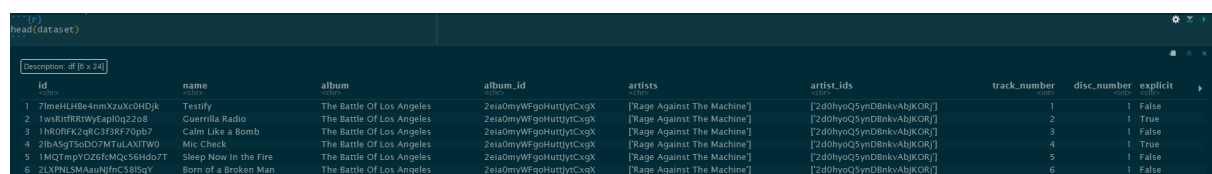
We performed an effective multivariate analysis of our Spotify music data using a variety of techniques. We explored the nature of the data using plots visualizing its nature in varying dimensions including 1D, 2D, and more. Furthermore, we created various models such as LDA, clustering methods, PCA, MDS, and even canonical correlation to obtain deeper insights from our data. Measures of association were also performed to understand more about the sparse categorical variables we had and whether any relationship between them was present or not.

Moreover, outlier detection methods were employed for single variables, between each pair of numeric variables as well as a multivariate method of Mahalanobis distance for more extensive insight. A large portion of our data was considered as outliers, however, they would be deemed as significant outliers as variability in musical data is expected, and some will stray from typical ranges. Upon removing the outliers, we reran our aforementioned experiments. Some displayed an improvement in results such as LDA and PCA, while others had a very small and nearly negligible difference in quality such as the k-means clustering and MDS.

In conclusion, there is still a lot to be desired in terms of how much information and storytelling can be extracted from such a small amount of data. However, this is a strong start to encourage deeper and more insightful data mining and visualization techniques in future work.

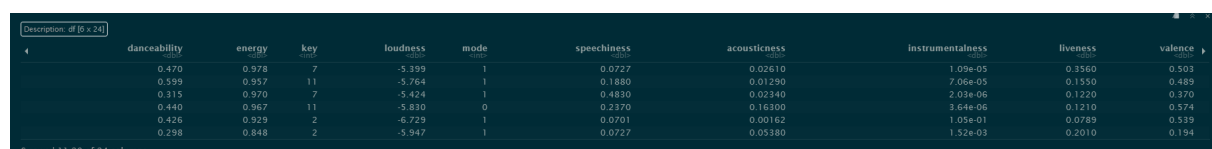
6 Appendix

6.1 Annotated Computer Output (Tables & Figures)



id	name	album	album_id	artists	artist_ids	track_number	disc_number	explicit
1	7lmeHLHBe4nmXzuCOHDJk	The Battle Of Los Angeles	2eia0myWfgohuttytCcxgX	[Rage Against The Machine]	[2id0hyoQ5ynD8nkVAbjKORJ]	1	1	False
2	1wsRitRRWYEploq2Zo8	The Battle Of Los Angeles	2eia0myWfgohuttytCcxgX	[Rage Against The Machine]	[2id0hyoQ5ynD8nkVAbjKORJ]	2	1	True
3	1hR0RfK2qG3FR70pb7	The Battle Of Los Angeles	2eia0myWfgohuttytCcxgX	[Rage Against The Machine]	[2id0hyoQ5ynD8nkVAbjKORJ]	3	1	False
4	2lBASgT5c0D7MTLAXITW0	Mic Check	2eia0myWfgohuttytCcxgX	[Rage Against The Machine]	[2id0hyoQ5ynD8nkVAbjKORJ]	4	1	True
5	1MqTmptQ258Gmc35hdo7T	Sleep Now in the Fire	2eia0myWfgohuttytCcxgX	[Rage Against The Machine]	[2id0hyoQ5ynD8nkVAbjKORJ]	5	1	False
6	2LXPNL3MAauNjfnC58ISqY	Born of a Broken Man	2eia0myWfgohuttytCcxgX	[Rage Against The Machine]	[2id0hyoQ5ynD8nkVAbjKORJ]	6	1	False

Figure 17: Dataset Observations Example



danceability	energy	key	loudness	mode	speechiness	acousticness	instrumentalness	liveness	valence
0.470	0.978	7	-5.399	1	0.0227	0.02610	1.09e-05	0.3560	0.503
0.399	0.957	11	-5.764	1	0.1800	0.01290	7.06e-05	0.1150	0.489
0.315	0.970	7	-5.424	1	0.4830	0.02340	2.03e-06	0.1220	0.370
0.440	0.967	11	-5.830	0	0.2370	0.16300	3.64e-06	0.1210	0.574
0.426	0.929	2	-6.729	1	0.0701	0.00162	1.05e-01	0.0789	0.539
0.298	0.848	2	-5.947	1	0.0727	0.05380	1.52e-03	0.2010	0.194

Figure 18: Dataset Observations Example

Description: df [8 x 24]									
	speechiness	acousticness	instrumentalness	liveness	valence	tempo	duration_ms	time_signature	year
	0.0727	0.02610	1.09e-05	0.3560	0.503	117.906	210133	4	1999
	0.1880	0.01290	7.06e-05	0.1550	0.469	103.680	206200	4	1999
	0.4830	0.02340	2.03e-06	0.1220	0.370	149.749	298893	4	1999
	0.2370	0.16300	3.64e-06	0.1210	0.574	96.752	213640	4	1999
	0.0701	0.00162	1.05e-01	0.0789	0.539	127.059	205600	4	1999
	0.0727	0.05380	1.52e-03	0.2010	0.194	148.282	280960	4	1999

6 rows | 16-25 of 24 columns

Figure 19: Dataset Observations Example

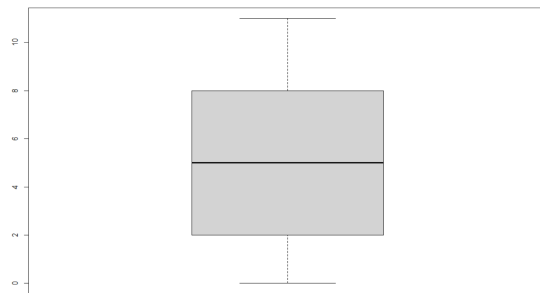


Figure 20: Key Box Plot

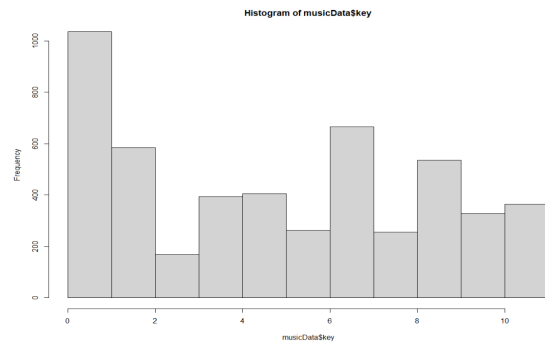


Figure 21: Key Histogram

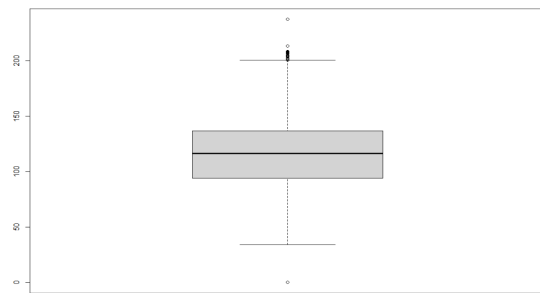


Figure 22: Tempo Box Plot

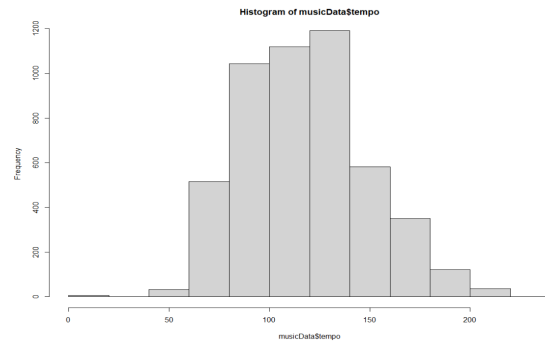


Figure 23: Tempo Histogram

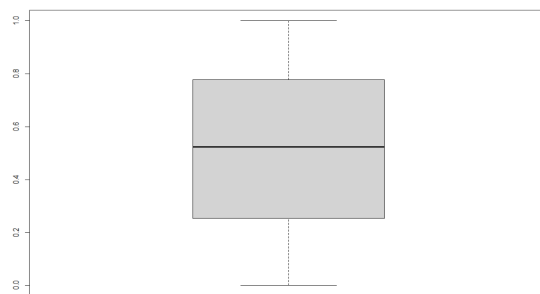


Figure 24: Energy Box Plot

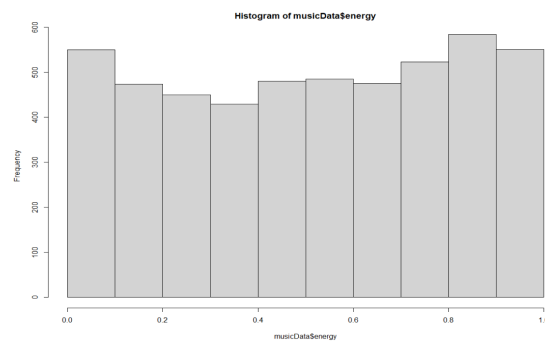


Figure 25: Energy Histogram

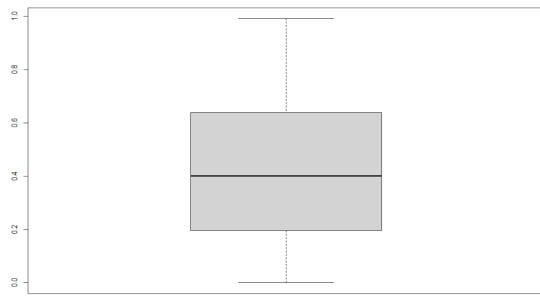


Figure 26: Valence Box Plot

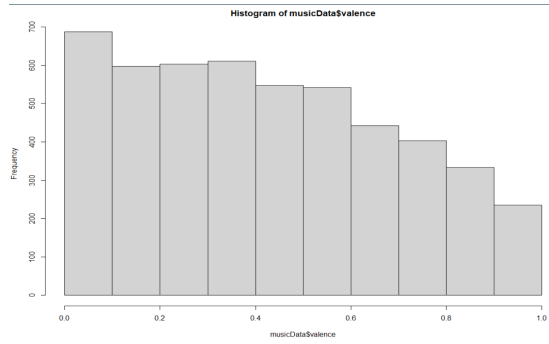


Figure 27: Valence Histogram



Figure 28: Liveness Box Plot

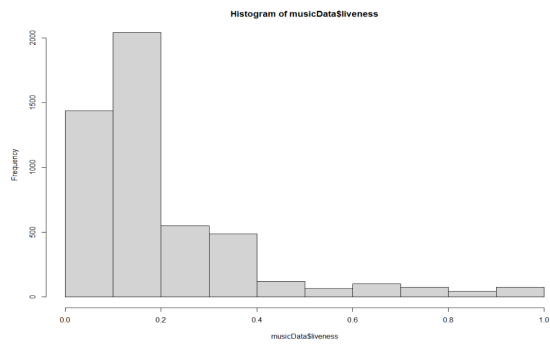


Figure 29: Liveness Histogram

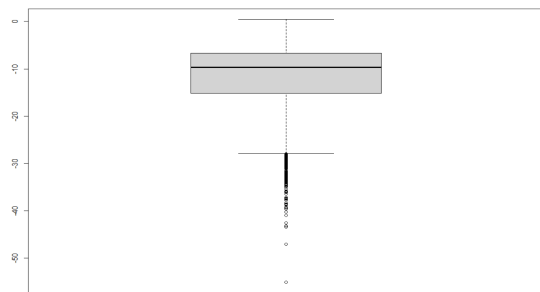


Figure 30: Loudness Box Plot

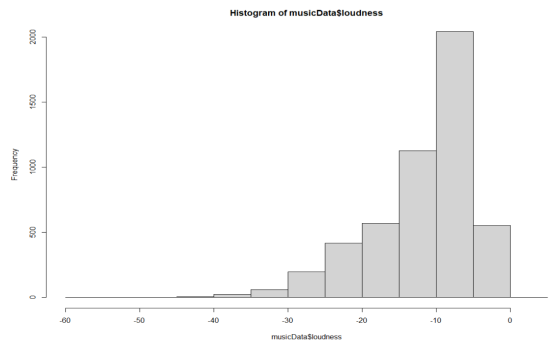


Figure 31: Loudness Histogram

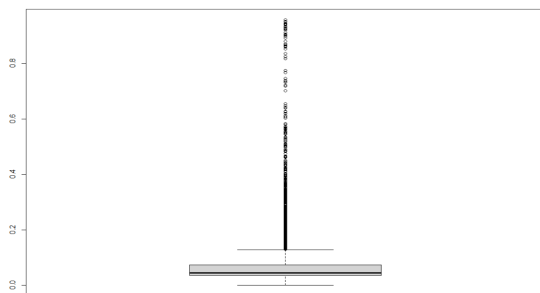


Figure 32: Speechiness Box Plot

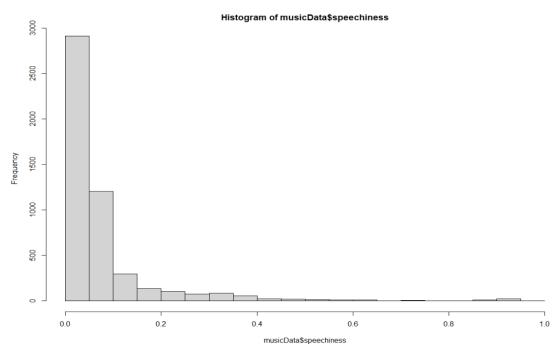


Figure 33: Speechiness Histogram

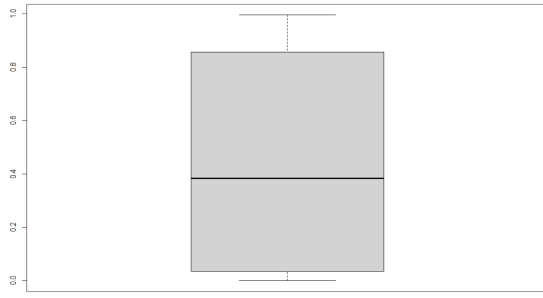


Figure 34: Acousticness Box Plot

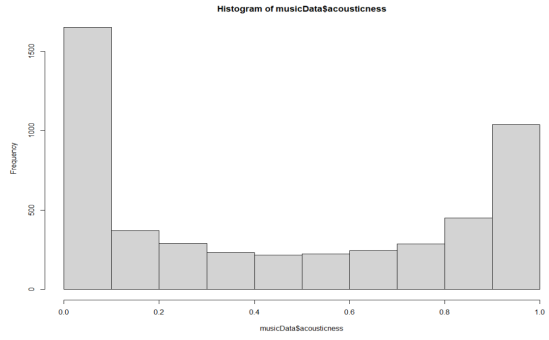


Figure 35: Acousticness Histogram

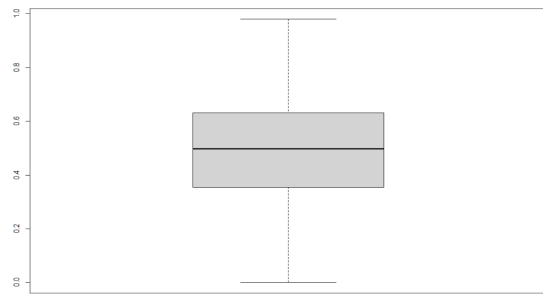


Figure 36: Danceability Box Plot

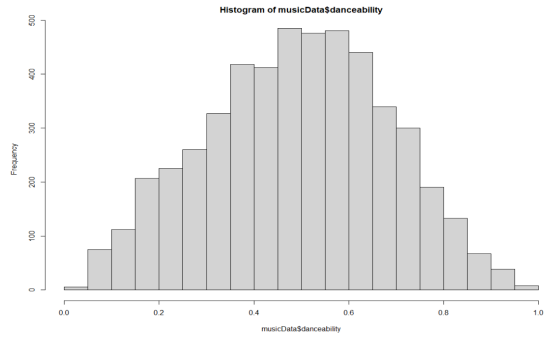


Figure 37: Danceability Histogram

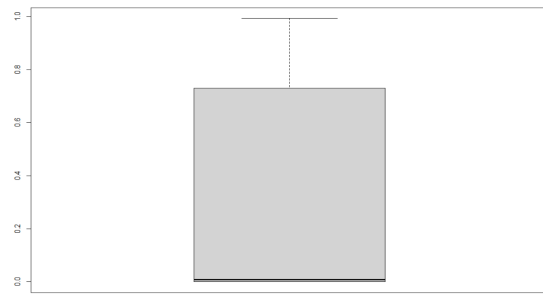


Figure 38: Instrumentalness Box Plot

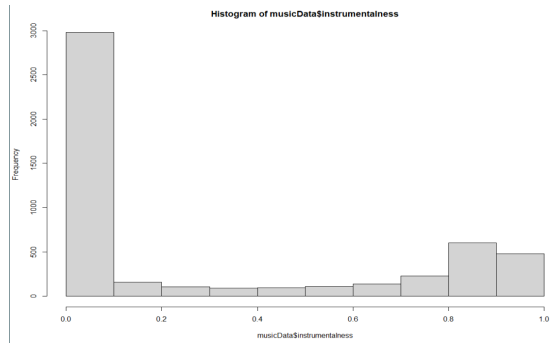


Figure 39: Instrumentalness Histogram

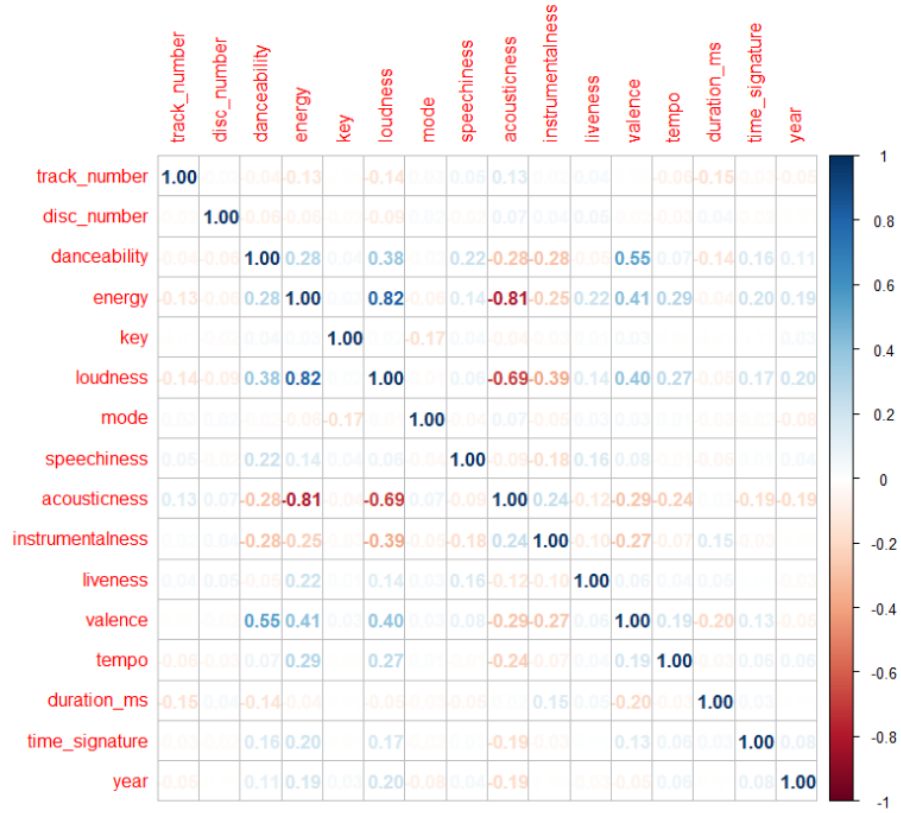


Figure 40: Dataset Observations Example

Description: df [50,000 x 11]											
key	danceability	energy	loudness	mode	speechiness	acousticness	instrumentalness	liveness	valence	tempo	
11	0.6270	0.959000	-2.061	0	0.2810	2.34e-02	1.62e-03	0.7220	0.2050	140.018	
10	0.6150	0.667000	-8.309	1	0.0298	1.23e-01	5.58e-05	0.1880	0.9600	125.922	
0	0.5300	0.633000	-9.227	1	0.0348	5.94e-01	6.90e-04	0.1170	0.7370	119.025	
9	0.4680	0.424000	-10.252	1	0.1850	7.68e-01	0.00e+00	0.0964	0.3040	135.445	
10	0.7190	0.491000	-9.723	1	0.1250	9.03e-01	0.00e+00	0.2680	0.8590	126.388	
8	0.3480	0.988000	-3.599	1	0.0705	5.63e-05	2.44e-06	0.2860	0.4330	165.057	
0	0.4670	0.164000	-16.338	1	0.0417	9.57e-01	9.31e-01	0.0663	0.3740	118.591	
10	0.7420	0.228000	-21.630	0	0.1990	3.11e-01	6.69e-01	0.0603	0.3200	122.123	
4	0.6850	0.858000	-5.807	0	0.0495	6.49e-06	7.83e-01	0.0967	0.2510	110.002	
7	0.5810	0.661000	-6.147	1	0.0312	1.51e-03	1.62e-03	0.0865	0.4530	79.975	
8	0.6930	0.693000	-10.764	1	0.0851	4.94e-02	0.00e+00	0.7440	0.6850	136.884	
8	0.6230	0.917000	-9.687	1	0.0390	6.53e-02	6.71e-01	0.3550	0.3340	131.997	
9	0.3500	0.578000	-7.546	0	0.0416	8.49e-02	0.00e+00	0.2110	0.1890	150.056	
8	0.6080	0.403000	-10.685	1	0.0597	8.72e-01	4.73e-05	0.1260	0.3600	126.419	
0	0.2380	0.028700	-33.691	1	0.0533	7.42e-01	1.62e-02	0.0575	0.4040	123.981	
5	0.5320	0.761000	-4.973	1	0.0307	1.90e-01	0.00e+00	0.1560	0.8270	134.844	
9	0.5590	0.925000	-7.470	1	0.0381	1.21e-02	2.68e-02	0.9690	0.4530	132.634	
3	0.1460	0.750000	-8.103	0	0.0611	3.83e-01	9.07e-01	0.1590	0.1530	82.916	
0	0.6170	0.458000	-12.095	0	0.0431	9.03e-02	0.00e+00	0.0820	0.9430	134.431	
5	0.3410	0.965000	-3.369	1	0.0998	1.44e-06	0.00e+00	0.2410	0.0314	119.815	
10	0.5350	0.477000	-8.498	1	0.0418	6.42e-01	2.46e-06	0.0798	0.6960	103.768	
11	0.6830	0.631000	-5.873	0	0.0863	1.54e-01	1.11e-05	0.3110	0.7330	120.026	
4	0.6340	0.188000	-20.458	1	0.9110	8.60e-01	0.00e+00	0.3750	0.4440	74.239	
0	0.3820	0.140000	-15.760	1	0.0623	6.80e-01	0.00e+00	0.5590	0.1800	94.204	
10	0.5670	0.295000	-9.832	1	0.0411	9.61e-01	0.00e+00	0.2770	0.9090	84.718	
7	0.3210	0.328000	-10.578	1	0.0294	8.57e-01	2.99e-02	0.2210	0.3210	75.003	
7	0.2340	0.204000	-17.302	1	0.0349	9.16e-01	8.35e-01	0.1000	0.1860	151.059	

Figure 41: New Variables for LDA

Description: df [50,000 x 11]										
key <int>	danceability <float>	energy <float>	loudness <float>	mode <int>	speechiness <float>	acousticness <float>	instrumentalness <float>	liveness <float>	valence <float>	tempo <float>
11	0.705961961	1.521242291	1.3946490024	-1.4323815	1.6631007372	-1.090427637	-0.745891730	2.884002128	-8.292088e-01	0.7157932074
10	0.642910316	0.530881407	0.4993562648	0.6981241	-0.4683672259	-0.832302964	-0.750055105	-0.075664528	1.965353e+00	0.2602660699
0	0.196294502	0.415565413	0.3678135739	0.6981241	-0.4259415101	0.388346841	-0.748367078	-0.469178259	1.139939e+00	0.0373822257
9	-0.129472327	-0.293288781	0.2209385649	0.6981241	0.8485269934	0.839287534	-0.750203626	-0.583352666	-4.627695e-01	0.5680118794
10	1.189357901	-0.066048441	0.2967403988	0.6981241	0.3394184034	1.189155313	-0.750203626	0.367731226	1.591511e+00	0.2753253524
8	-0.759888771	1.619600050	1.1742648426	0.6981241	-0.1230218991	-1.150925477	-0.750197132	0.467495271	1.471189e-02	1.5249549551
0	-0.134726631	-1.175116966	-0.6511407565	0.6981241	-0.3673940223	1.329102424	1.727805938	-0.750180218	-2.026771e-01	0.0263834798
10	1.310206886	-0.958051567	-1.4094456808	-1.4323815	0.9673189977	-0.345079687	1.030449863	-0.783434999	-4.035470e-01	0.1374973699
4	1.010711575	1.178685958	0.8578745794	-1.4323815	-0.3012099056	-1.151054565	1.333879605	-0.581689932	-6.589441e-01	-0.2542055553
7	0.359177917	0.510531525	0.8091550642	0.6981241	-0.4564880255	-1.147158049	-0.745891730	-0.638222890	8.874002e-02	-1.2245597947
8	1.052746005	0.619064225	0.1475727068	0.6981241	0.0008611911	-1.023045694	-0.750203626	3.005935960	9.474663e-01	0.6145146852
8	0.684944746	1.378793123	0.3018989357	0.6981241	-0.3903039088	-0.981839045	1.035773192	0.849924108	-3.517273e-01	0.4563861153
9	-0.749480164	0.229024836	0.6086885886	-1.4323815	-0.3682425366	-0.931043426	-0.750203626	0.051811752	-8.884313e-01	1.0401817661
0	0.806130190	-0.264512385	0.1588928294	0.6981241	-0.2146614453	1.108815304	-0.750077729	-0.419296237	-2.554908e-01	0.2763271501
0	-1.337962178	-1.634006787	-3.1376988350	0.6981241	-0.2689663615	0.771905592	-0.707084663	-0.789953851	-1.438830e+00	0.1975406035
5	0.206803109	0.84966212	0.9773806842	0.6981241	-0.4607305971	-0.658664882	-0.750203626	-0.253022829	1.473066e+00	0.5485899292
9	0.348669309	1.405926298	0.6195788331	0.6981241	-0.3979405377	-1.119712866	-0.678871020	4.252986518	8.874002e-02	0.4771714435
3	-1.821358118	0.812388096	0.5288745593	-1.4323815	-0.2027822449	-0.158483539	1.663925992	-0.236395488	-1.021682e+00	-1.1295182715
0	0.653418924	-0.177972788	-0.0431498658	-1.4323815	-0.3555148218	-0.917048715	-0.750203626	-0.663163901	1.902429e+00	0.5352433977
5	-0.796763897	1.609425109	1.2072221617	0.6981241	0.1255927957	-1.151067653	-0.750203626	0.772329851	-1.471773e+00	0.0629119106
10	0.539242424	-0.113531498	0.4722729461	0.6981241	-0.3665455079	0.515335887	-0.750197078	-0.675352284	9.881818e-01	2.1296206950
11	0.895116894	0.408782120	0.8484172617	-1.4323815	0.0110433629	-0.751962956	-0.750174082	0.606056444	1.125134e+00	0.0697305986
4	0.742742086	-1.093717441	-1.2415066462	0.6981241	7.0087409315	1.077715946	-0.750203626	0.960773047	5.542736e-02	-1.4099246968
0	-0.581342445	-1.256516491	-0.5683175807	0.6981241	-0.1926000731	0.611225575	-0.750203626	1.980583280	-9.217439e-01	-0.7647346216
10	0.390703739	-0.730811227	0.2811214954	0.6981241	-0.3724851082	1.339468877	-0.750203626	0.417613248	1.776581e+00	-1.0712847370
7	-0.901854971	-0.618886880	0.1742231475	0.6981241	-0.4717612832	1.069941107	-0.670619860	0.107236221	-3.998456e-01	-1.3852352294
7	-1.358979392	-1.039451091	-0.7892749113	0.6981241	-0.4250929958	1.222846284	1.472286155	-0.563399857	-8.995355e-01	1.0725947912

Figure 42: Scaling Numeric Features

# finding mean of each predictor					
```{r}					
apply(music_num[, -1], 2, mean)					
danceability	energy	loudness	mode	speechiness	
-1.480759e-17	-1.769217e-16	7.609519e-17	-1.403823e-17	-1.665928e-17	
acousticness	instrumentalness	liveness	valence	tempo	
9.782539e-18	-4.389310e-17	4.513853e-17	-8.695493e-17	-1.293016e-16	
# finding standard deviation of each predictor					
```{r}					
apply(music_num[, -1], 2, sd)					
danceability	energy	loudness	mode	speechiness	
1	1	1	1	1	
acousticness	instrumentalness	liveness	valence	tempo	
1	1	1	1	1	
# create training and test sets					
```{r}					
set.seed(1)					
# 70% for training 30% for test					
sample <- sample(c(TRUE, FALSE), nrow(music_num), replace = TRUE, prob = c(0.7, 0.3))					
train <- music_num[sample, ]					
test <- music_num[!sample, ]					
train					
test					
```					

Figure 43: Feature Scaling and Train-Test Split

```
Call:
lda(train$key ~ ., data = train)

Prior probabilities of groups:
      0      1      2      3      4      5      6      7      8
0.12570399 0.07979073 0.11584093 0.03319134 0.08113439 0.08396466 0.05397524 0.12987793 0.05560479
      9     10     11
0.11015181 0.06266617 0.06809800

Group means:
      danceability      energy      loudness      mode      speechiness      acousticness      instrumentality
0      0.01703145 -0.093565583 -0.03201586 0.29353636 -0.09823243 0.08244485 -0.005757267
1      0.12973428 0.175445395 0.08882260 0.11263579 0.24987445 -0.25231981 -0.009381582
2      -0.07483587 -0.034906542 -0.01163919 0.22964419 -0.08748356 0.04502880 0.020535639
3      -0.28654885 -0.355095475 -0.31673684 0.08888388 -0.12350718 0.39289820 0.248191106
4      -0.05866124 0.051850226 -0.07265567 -0.38214285 -0.06366447 -0.03163917 -0.048082811
5      -0.02851295 -0.168075956 -0.12762132 -0.04976536 -0.05523291 0.19816643 0.059440564
6      0.05078285 0.196282552 0.10606564 -0.29829353 0.21353942 -0.21461888 -0.033562306
7      -0.00790835 -0.042271570 -0.02436849 0.29153187 -0.09483415 0.03456954 0.003273891
8      0.04536280 -0.009118008 -0.03347713 0.18658365 0.17732861 0.01654112 0.023745798
9      -0.02937450 0.037244578 0.03672240 -0.15839115 -0.07319541 -0.02517950 -0.047857200
10     0.05601083 -0.081141508 -0.08026326 -0.28645720 0.07478103 0.08988749 0.065236001
11     0.09168776 0.190428978 0.12418809 -0.53796271 0.13306129 -0.19663891 -0.089937160

      liveness      valence      tempo
0      -0.013885971 0.008427751 0.003460935
1      0.008517913 -0.058057545 0.010705697
2      -0.012970726 -0.022000816 0.010588179
3      -0.074632420 -0.204749942 -0.140603287
4      0.029101390 -0.023012717 0.034933880
5      -0.010258559 0.001382153 -0.043337552
6      0.016804800 0.023406472 -0.046151563
7      0.021822758 0.048664046 0.006684604
8      -0.025453663 -0.066937136 -0.024086736
9      -0.022423003 0.055329567 0.036559445
10     -0.019395038 -0.020769263 -0.056842506
11     0.024039951 0.048008233 0.053394357

Coefficients of linear discriminants:
      LD1      LD2      LD3      LD4      LD5      LD6
danceability -0.04305622 -0.37049904 0.2607327581 0.98272434 -0.40441504 -0.158097420
energy      0.23590122 -0.44430058 -0.0006041129 -0.41351532 0.33851699 0.916861759
loudness    -0.08140902 0.43535777 -0.1168509896 -0.10472333 -0.73576725 0.007339267
mode        -0.98925609 -0.31544083 -0.0725154234 -0.06470487 0.03063914 0.077936274
speechiness 0.13882396 -0.55209322 0.5166075276 -0.15854834 0.14288613 0.373620474
acousticness -0.05380213 0.61331877 0.4924653397 0.02788668 -0.24432488 1.079782551
instrumentality -0.10033889 -0.05624498 0.4040083790 -0.10826830 0.13346261 -0.507436104
liveness    -0.03545891 -0.11904005 -0.0746867455 0.24728139 -0.06814184 -0.524059010
valence     -0.01688250 0.45436453 -0.2803678091 0.08940298 1.04701340 0.094656950
tempo       -0.02094438 0.05547593 -0.1759653451 0.17591476 -0.39822558 0.216644748

      LD7      LD8      LD9      LD10
danceability 0.16808979 0.044704554 -0.30191368 -0.5188239
energy      0.42373842 0.909572505 -0.73123031 -1.6244875
loudness    -1.10185884 -0.740194483 -0.39747361 1.0112810
mode        -0.02978845 0.025334092 -0.01048862 -0.0262324
speechiness -0.23963999 0.009374529 0.17146997 0.5481275
acousticness -0.29507450 0.498360123 -0.61508377 -0.6430607
instrumentality -0.07334918 -0.021112152 -0.80992349 0.4203133
liveness    -0.49100257 0.706769433 0.15811459 0.0263149
valence     -0.13328870 -0.122127877 -0.01409618 0.5611427
tempo       0.57043755 0.477278281 -0.09683527 0.5108648

Proportion of trace:
      LD1      LD2      LD3      LD4      LD5      LD6      LD7      LD8      LD9      LD10
0.6631 0.2296 0.0665 0.0223 0.0101 0.0044 0.0022 0.0014 0.0004 0.0000
```

Figure 44: LDA Model

```
##{r}
head(predicted$class)
##
[1] 0 7 0 7 7 4
Levels: 0 1 2 3 4 5 6 7 8 9 10 11
```

Figure 45: Classes for Key Label

```
##{r}
head(predicted$posterior)
##
      0      1      2      3      4      5      6      7
4 0.15158530 0.07927503 0.13754239 0.03738247 0.06142439 0.08669892 0.03915476 0.13749041
6 0.13112696 0.09519697 0.14675968 0.02101267 0.07470584 0.06036773 0.05070455 0.15110307
7 0.17099703 0.05400948 0.14299366 0.06304616 0.04835146 0.10683186 0.02604046 0.15993725
15 0.14277083 0.09792789 0.13242643 0.04135841 0.04887567 0.07947084 0.03976427 0.15732968
17 0.15038009 0.09692281 0.12970251 0.01712459 0.06925217 0.06973242 0.04215737 0.18440721
18 0.06017875 0.05800842 0.08740351 0.04179379 0.14316136 0.07712601 0.09654853 0.06888463
      8      9      10      11
4 0.07821533 0.09704859 0.05370199 0.04048041
6 0.05838118 0.11898400 0.03718850 0.05446885
7 0.06114645 0.08497037 0.05697870 0.02469713
15 0.07104259 0.10071282 0.04732294 0.04099763
17 0.05155596 0.09563673 0.04272649 0.05040164
18 0.03931957 0.13274404 0.08079591 0.11403548
```

Figure 46: Posterior Probabilities for LDA

```

[[r]]
mean(predicted$class==test$key)
[[r]]
[1] 0.1582451

```

Figure 47: LDA Model Accuracy

```

[[r]]
key_label <- music_num$key
music_num$key <- NULL
music_num
[[r]]

```

danceability <dbl>	energy <dbl>	loudness <dbl>	mode <dbl>	speechiness <dbl>	acousticness <dbl>
0.705961961	1.521242291	1.3946490024	-1.4323815	1.6631007372	-1.090427637
0.642910316	0.530881407	0.4993562648	0.6981241	-0.4683672259	-0.832302964
0.196294502	0.415565413	0.3678135739	0.6981241	-0.4259415101	0.388346841
-0.129472327	-0.293288781	0.2209385649	0.6981241	0.8485269934	0.839287534
1.189357901	-0.066048441	0.2967403988	0.6981241	0.3394184034	1.189155313
-0.759988771	1.619600050	1.1742648426	0.6981241	-0.1230218991	-1.150925477
-0.134726631	-1.175116966	-0.6511407565	0.6981241	-0.3673940223	1.329102424
1.310206886	-0.958051567	-1.4094456808	-1.4323815	0.9673189977	-0.345079687
1.010711575	1.178685958	0.8578745794	-1.4323815	-0.3012099056	-1.151054565
0.359177917	0.510531525	0.8091550642	0.6981241	-0.4564880255	-1.147158049

Figure 48: Removing Label from Dataset

```

[[r]]
# Dissimilarity matrix
d <- dist(music_num_sample_for_clustering, method = "euclidean")
# Hierarchical clustering using complete linkage
hcl <- hclust(d, method = "complete")
# Plot the obtained dendrogram
plot(hcl, cex = 0.6, hang = -1)
[[r]]

```

Figure 49: Hierarchical Clustering Model

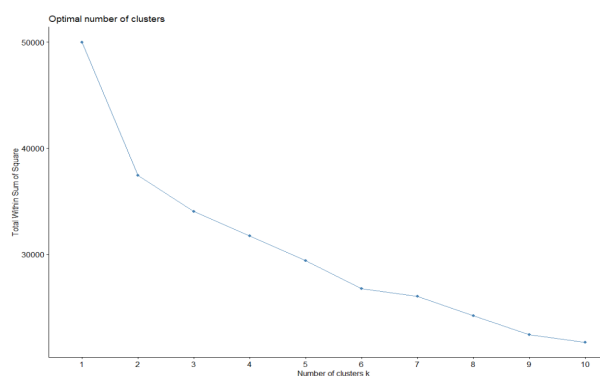


Figure 50: Elbow Method for K-Means

```

[[r]]
km$iter
[[r]]
[1] 4

```

Figure 51: K-Means Number Iterations

```
# Testing normality using mshapiro test
library(r)
mshapiro.test(t(hypothesisData))

Shapiro-Wilk normality test

data:  z
W = 0.94385, p-value < 2.2e-16
```

Figure 52: Normality Check

```
# Checking determinant of var-cov matrix is positive
library(r)
det(cov(hypothesisData))

[1] 4.055918e-05
```

Figure 53: Determinant Check

```
Call:
lda(train2$key ~ ., data = train2)

Prior probabilities of groups:
 0 1 2 3 4 5 6 7 8 9 10
0.12697368 0.07434211 0.12006579 0.03421053 0.08289474 0.08322368 0.05460526 0.13289474 0.04802632 0.11217105 0.06250000
0.06809211

Group means:
  danceability energy loudness mode speechiness acousticness instrumentality liveness valence
0 -0.06260739 -0.13505093 -0.10836349 0.34000111 -0.119282204 0.18473534 0.060801427 0.018428907 -0.01651347
1 -0.13459498 0.27834608 0.14977052 0.07009072 0.284359986 -0.29031313 -0.012636997 -0.091655136 0.04293547
2 -0.09105190 -0.15914895 -0.09272059 0.23750948 -0.117054997 0.14153549 -0.005129926 -0.006723956 -0.11261430
3 -0.29317449 -0.32191504 -0.43503847 0.07604405 -0.237347096 0.48413628 0.336355898 -0.179503297 -0.32341128
4 -0.05305385 0.09728840 0.06336627 -0.32681503 -0.001973215 -0.08192521 -0.017245153 0.076972012 -0.00750521
5 -0.02635170 -0.09997796 -0.05485487 -0.05456908 -0.059336158 0.15723534 0.026552831 0.017051671 0.06869780
6 -0.15058216 0.22604300 0.15573998 -0.51808683 0.146118255 -0.29188602 -0.089239668 -0.014467366 0.13217246
7 -0.00134078 -0.02647687 -0.01015493 0.31966253 -0.042962170 0.01816756 -0.016896886 -0.029245419 0.06477046
8 -0.03992562 0.03014188 0.04023146 0.19975416 0.084711233 -0.04493305 -0.070898754 0.029394623 -0.13346079
9 -0.01890863 0.03270310 0.01094199 -0.13133949 -0.134820119 -0.04753079 -0.006969543 0.090046611 0.05263821
10 -0.12054904 -0.06243486 -0.07015688 -0.38552976 0.027879602 0.04463054 0.041770393 0.023435249 0.03140983
11 0.11859388 0.21429615 0.15942530 -0.49032221 0.138856391 -0.23446241 -0.199933448 -0.050439092 0.05707722

tempo
0 0.005213537
1 0.063892054
2 -0.037777400
3 -0.256686089
4 0.053973181
5 -0.008777273
6 -0.103072861
7 0.014334791
8 -0.004193621
9 -0.027757956
10 -0.082539976
11 0.123225925

Coefficients of linear discriminants:
  LD1 LD2 LD3 LD4 LD5 LD6 LD7 LD8 LD9
danceability 0.04330627 -0.1802355 -0.23943560 0.269438871 -0.43937318 0.25828747 -0.76206079 0.72988674 -0.11598799
energy 0.18059113 -0.8263924 0.32295888 0.021971560 -0.80927218 -1.28141816 -0.07750744 0.91642874 -1.82903515
loudness -0.26490005 0.6338264 -0.09346871 -0.232645984 0.06529876 0.27413989 -0.01974900 -1.56245702 0.46288466
mode -0.93810287 -0.4663287 0.06120250 0.068951113 0.08677938 0.02069061 -0.08436238 0.02686752 -0.01966404
speechiness 0.13748255 -0.4275243 -0.53655832 0.076078298 -0.31847800 0.02710715 -0.02582459 -0.59849847 0.32797627
acousticness -0.30247624 0.4324263 0.03886718 -0.046104557 -1.33443633 -0.61442372 -0.10074136 -0.23869675 -1.17536775
instrumentality -0.16023113 0.1413327 -0.39171229 0.348590106 0.12363371 -0.61311473 0.01159061 -0.19981755 0.45299617
liveness -0.03457143 0.3108715 0.38076976 0.008863652 0.02963188 -0.25953256 -0.75911385 -0.07591333 0.34516923
valence 0.03107936 0.2794021 0.45211917 0.842339946 0.15181782 0.19344107 0.61044048 -0.46360755 0.25651235
tempo -0.02078166 -0.0464572 0.26082968 -0.320435032 -0.59263788 -0.07581930 0.30009324 0.33688179 0.62242010

LD10
danceability -0.498156575
energy 0.405573855
loudness -1.184605227
mode 0.005338641
speechiness 0.423120679
acousticness -0.044287309
instrumentality -0.525566014
liveness 0.22992745
valence 0.394912773
tempo -0.102744725

Proportion of trace:
  LD1 LD2 LD3 LD4 LD5 LD6 LD7 LD8 LD9 LD10
0.6591 0.1957 0.0507 0.0352 0.0274 0.0153 0.0099 0.0052 0.0012 0.0002
```

Figure 54: LDA Summary after outlier removal


```

```{r}
predicted <- predict(model, test)
names(predicted)
mean(predicted$class==test$key)
```

[1] "class"      "posterior" "x"
[1] 0.1499673

```

Figure 55: LDA MSE after outlier removal

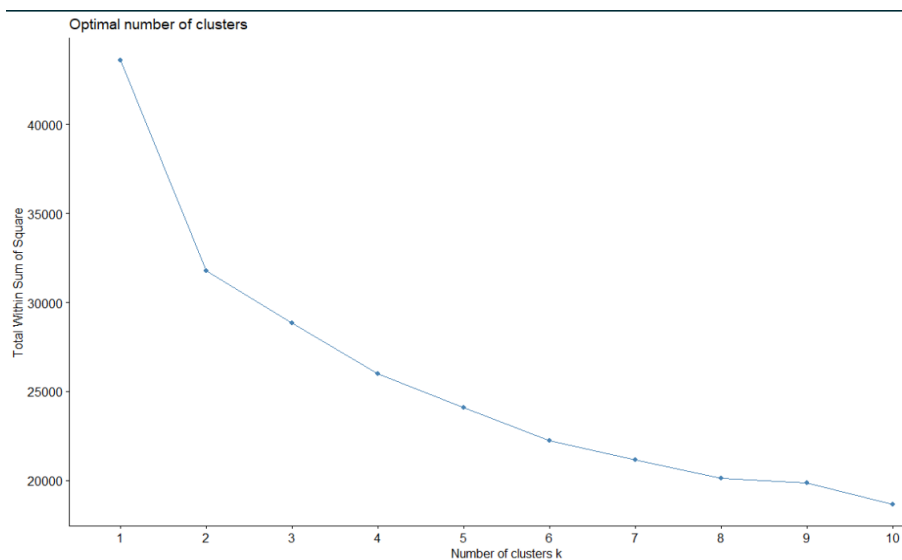


Figure 56: Elbow method after outlier removal

```

```{r}
mds2 <- cmdscale(d2, eig = T, k=2)
mds2$GOF
```

[1] 0.4792074 0.4792074

```

Figure 57: MDS after outlier removal

References

- [1] G. Andrienko, N. Andrienko, S. Drucker, J.-D. Fekete, D. Fisher, S. Idreos, T. Kraska, G. Li, K.-L. Ma, J. Mackinlay, *et al.*, “Big data visualization and analytics: Future research challenges and emerging applications,” in *BigVis 2020-3rd International Workshop on Big Data Visual Exploration and Analytics*, 2020.
- [2] I. Frades and R. Matthiesen, “Overview on techniques in cluster analysis,” *Bioinformatics methods in clinical research*, pp. 81–107, 2010.