Middle East Technical University

Department of Statistics

Multivariate Analysis on Spotify Data Report

PROJECT REPORT SUBMITTED
IN FULFILMENT OF THE REQUIREMENTS FOR COURSE
STAT 467 – MULTIVARIATE ANALYSIS
DEPARTMENT OF STATISTICS OF
MIDDLE EAST TECHNICAL UNIVERSITY

BY

Alper Tunahan ÖZTÜRK	*2290856*
Dayanch AKMYRADOV	*2347292*
Kubilay TAŞYÜREK	*2218329*

February 2022

Contents

Abstract

Spotify is an online audio service that allows users to have nearly unlimited music at their disposal. No matter your tastes or preferences, you'll find something you like on Spotify. You can access Spotify through your browser or mobile app. Many songs become popular on the platform, but it is interesting to see what does affect the performance of a song. This research applies multivariate analysis to a dataset that includes many factors about the songs, mainly examining these factors to find valuable insights about the popularity of songs. Furthermore, another goal is to analyze a user's listening profile so that Spotify can recommend and acquire related tunes on their platform, thus improving the user experience. Through this analysis, it was concluded that there are a high number of false positives and that it is hard to predict the performance of a song, but it can be found what factors decrease the chances of a song being popular.

1 Introduction

With literally millions of songs on Spotify, some songs are bound to become more popular than others. According to various studies and publications, music can help with academic accomplishment, motivation, and the development of creativity [1]. We used multivariate analysis on a dataset that contains a variety of music-related variables, with the goal of uncovering useful information regarding song popularity and to assess a customer's listening profile in order for Spotify to recommend and acquire relevant tracks on their platform. This can be used to predict the popularity of a song for a given user, which is useful for target marketing purposes. We hope that this study will spark interest regarding the statistical and machine learning methods used in such analyses.

1.1 Data Description

ATTRIBUTE	TYPE	DEFINITION
year	Numerical	Release date of the song
valence	Numerical	Measure from 0.0 to 1.0 describing the musical positiveness conveyed by a track.
acousticness	Numerical	Measure of how acoustic the track is and ranges from 0.0 to 1.0
danceability	Numerical	Values range from 0.0 being least danceable and 1.0 being most danceable.
duration_ms	Numerical	The duration of the track in milliseconds(ms).
energy	Numerical	Measure from 0.0 to 1.0 and represents the energy of the song.
explicit	Categorical	1 for explicit 0 for not explicit.
insturmentalness	Numerical	Measure whether a track contains vocals. Values ranges from 0.0 to 1.0
key	Numerical	Estimated overall key of the track. If key is not detected, the value is -1.
liveness	Numerical	Detects the presence of an audience in the recording.
loudness	Numerical	Overall loudness of a track in decibels (dB). Values typical range between -60 and 0 dB.

mode	Categorical	Mode indicates the modality (major or minor) of a track. Major is represented by 1 and minor is represented by 0.
popularity	Numerical	Measure the popularity from 0 to 100 based on play number of the track.
speechiness	Numerical	Detects the presence of spoken words in a track.
tempo	Numerical	Overall estimated tempo of a track in beats per minute (BPM).

Table 1: Data Summary

The dataset includes 15 variables and 170653 observations. Out of 15 variables 2 of them are categorical and remaining variables are numeric. All of the variables describe the attributes of songs on Spotify as can seen on the above table. For example, the variable popularity measures the popularity of a song from 0 to 100 and danceability‘s range is between 0 to 1 and it measures the danceability of a track. One of the categorical variable explicit is describe the songs that has curse words or language. The dataset obtained from website on internet [2].

1.2 Research questions

The study was conducted to understand the affects of songs‘ attributes to its popularity. To achieve this purpose, some multivariate techniques used to find an answer to below questions.

- Using factors available, how well songs‘s popularity can be predicted?
- How to suggest someone similar type of songs according to one‘s favourite song.

1.3 Aim of the study

There are millions of different songs in the world today, and as a result, there are countless distinct forms of music in the globe. Most of these songs are unlikely to be popular, and only a handful are likely to be popular. Furthermore, each style of music seeks to target a specific set of individuals. As previously stated, there are several song characteristics in our data set. The mentioned research questions are attempted to be answered in this study with the use of these aspects.

2 Methodology

Hotelling T^2 The multivariate shape of the Student‘s t take a look at is the Hotelling T^2 take a look at. When Student’s t checks the equality of samples or the equality of a pattern imply to a particular imply, Hotelling T^2 checks the equality of multivariate pattern imply vectors or the equality of 1 multivariate pattern’s imply vector to a particular imply vector.

MANOVA OneWay MANOVA is a multivariable version of OneWay ANOVA. It compares the mean vectors of multiple responses with a categorical variable at more than two levels to see if they are the same. The sample must meet certain assumptions in order to use MANOVA. Multivariable normal matrices and common covariance matrices are essentially them.

Principal Component Analysis (PCA) Principal Component Analysis (PCA) is a multivariate data analysis statistical approach that is especially effective when working with three or more dimensional data. Due to the increased number of independent variables in multivariate analysis, the model's variance increases. As a result, the PCA approach allows the dataset's dimension to be reduced for further analysis. PCA is also an unsupervised approach for grouping variables in a dataset, and scree plots are used to generate a large number of factors or principal components.

Principal Component Regression The Essential Elements Regression is a method for assessing multicollinear data in multiple regression models. When multicollinearity exists, least squares estimates are unbiased, but their variances are large, therefore they may be far from the true value. By incorporating a degree of bias into the regression results, principal components regression reduces standard errors. The goal is for the final result to be more accurate estimates.

Factor Analysis Factor analysis is another dimesion reduction technique developed primarily in the field of psychology. The goal of this method is to define covariance relations between (p) in terms of a few (m) underlying and unobservable linear combinations known as factors. This approach works by grouping variables based on their correlations. Variables inside a group are associated between themselves, but not with variables from other groups. FA can be seen of as a progression of PCA.

Cannonical Correlation Analysis CCA is a technique for determining and quantifying relationships between two sets of data. CCA is a type of multivariate statistical analysis approach that examines several measurements on the same item at the same time (similar experimental units). CCA selects a collection of canonical variates, which are orthogonal linear combinations of variables within each set that best explain variability both within and across sets. CCA is a method for examining and measuring the relationship between two sets of variables, each of which contains two or more indicators.

Logisctic Regression Logistic regression is the technique of modeling the chance of a discrete final results given a given enter variable. The maximum not unusualplace logistic regression fashions binary outcomes. Something which can take values: true/false, yes/no, etc. Nominal logistic regression can version eventualities with a couple of feasible discrete final results. Logistic regression is a beneficial analytical method for class troubles whilst you are attempting to decide whether or not a brand new pattern is the quality in shape for a category.

Cluster Analysis Clustering is a phrase used to describe a set of methods for discovering subgroups of observations in a data set. When we cluster observations, we want observations in the same group to be similar and observations in other groups to be unique. Because there is no response variable, this is an unsupervised technique. Its goal is to find correlations between the n observations without using a response variable to train it. Clustering allows us to identify and categorize data that are similar.

Decision Tree Decision Trees are a widely used Data Mining approach that use a tree-like structure to offer outcomes depending on input decisions. The fact that decision trees may be utilized for both regression and classification is a significant feature. This kind of classification approach can handle both heterogeneous and missing data. Decision Trees can also generate rules that are easy to grasp. Furthermore, classifications do not need a large number of computations.

Random Forest Random Forest is a well-known supervised machine learning method. In Machine Learning, it can be applied to both classification and regression issues. It is based on ensemble learning, which is the process of combining numerous classifiers to solve a complicated issue and enhance the model's performance. Random Forest is a classifier that contains numerous number of decision trees on various subsets of a given dataset and takes the average to enhance the prediction accuracy of that dataset. Rather of depending on a single decision tree, the random forest takes each tree's forecast and predicts the final output based on the majority votes of predictions.

3 Results and Findings

3.1 Hotelling T^2

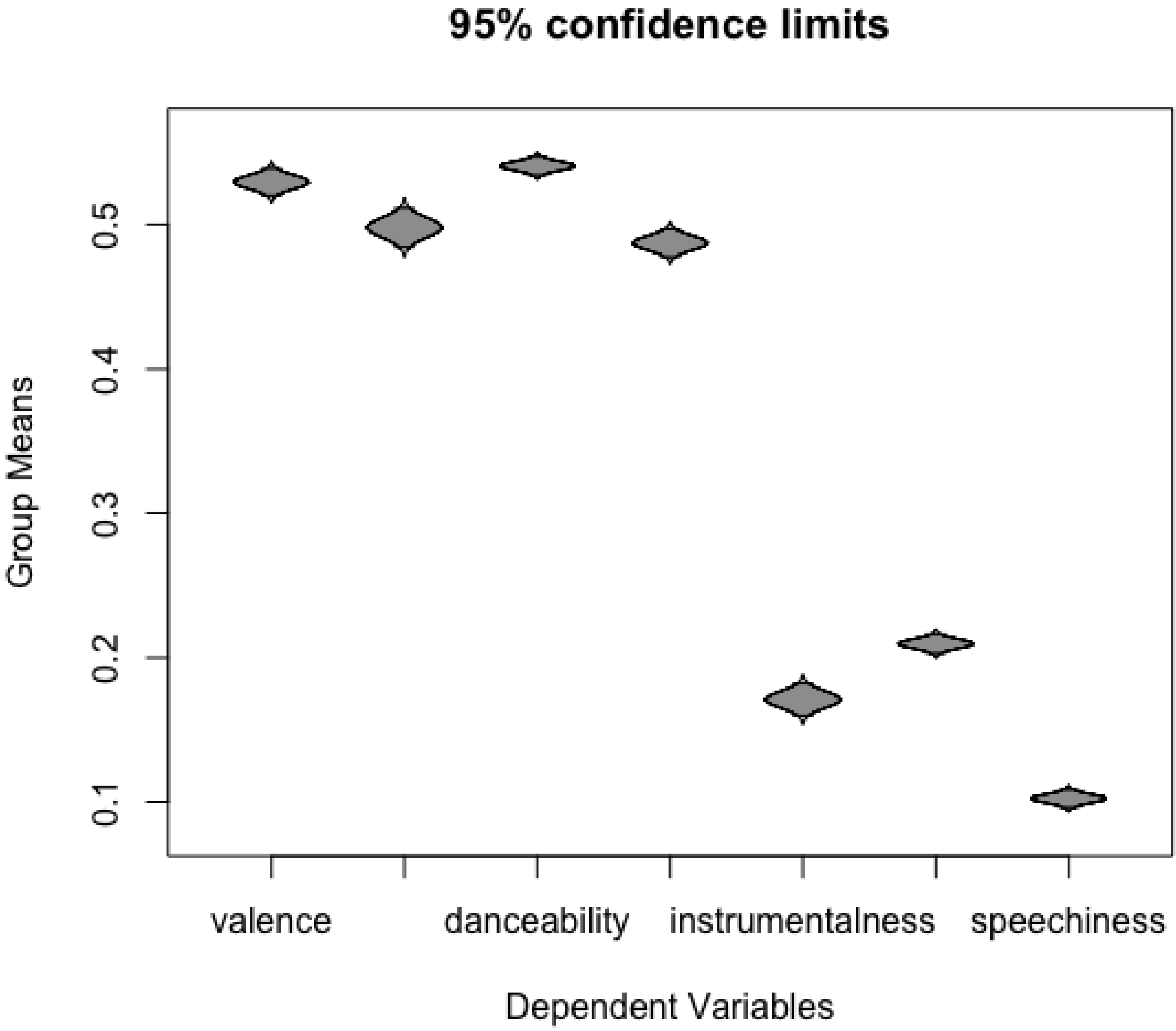


Figure 3.1: Error Bars for Group Means

Hotelling T^2	Dof	Dof2	p-value
14935	7	2993	<2.2e-16

Table 2: Hotelling T^2

The results for the one-sample multivariate t test indicated that the dependent variables means together were statistically significantly different from zero. The Hotelling T^2 value was statistically significant. Therefore, the null hypothesis of no joint mean difference was rejected.

3.2 MANOVA

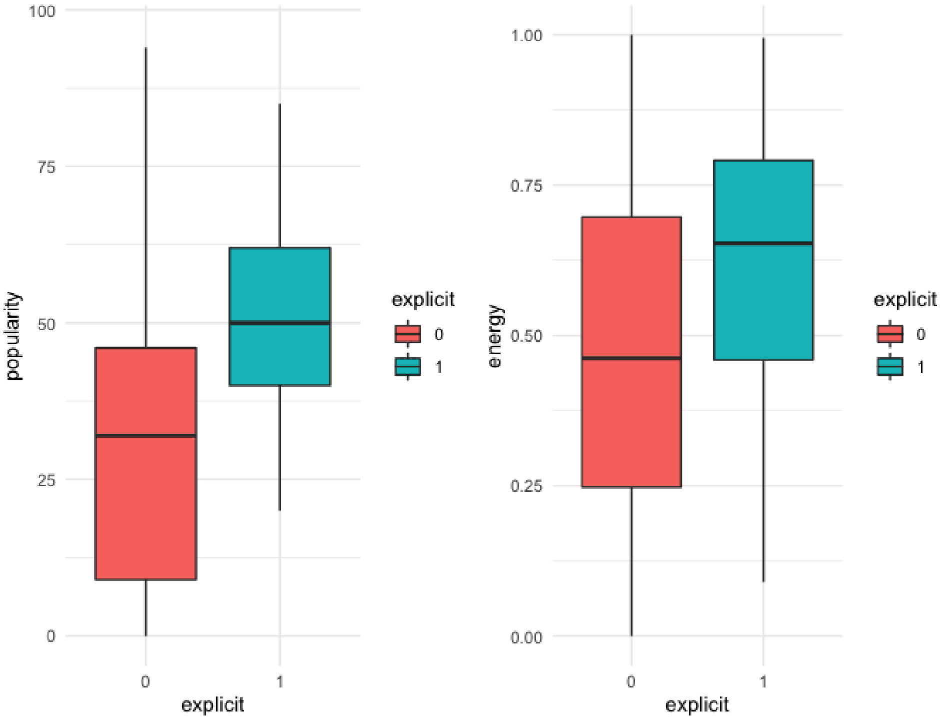


Figure 3.2: Boxplot of Explicit for Popularity and Energy

	Df	Pillai	approx F	num Df	den Df	Pr(>F)
(Intercept)	1	0.79	5622.58	2	2997	0.0000
explicit	1	0.03	49.28	2	2997	0.0000
Residuals	2998					

Table 3: MANOVA of Explicit for Popularity and Energy

Parameter	Eta2_partial	CI	CI_low	CI_high
explicit	0.04	0.95	0.03	1.00

Table 4: Partial Eta Squared

Explicit songs are better for popularity as well as for the energy, but popularity was more affected by explicit songs. The Pillai's Trace test statistics was statistically significant [Pillai's Trace = 0.041, F(2, 2997) = 64.65, $p < 0.001$] and indicates that explicit has a statistically significant association with both combined popularity and energy. The measure of effect size (Partial Eta Squared) was 0.04 and suggested that there is a medium effect of explicit content on popularity and energy.

MANOVA Assumptions

Assumptions of Multivariate Normality:

	explicit	variable	statistic	p
1	0	energy	0.96	0.00
2	0	popularity	0.95	0.00
3	1	energy	0.95	0.00
4	1	popularity	0.88	0.00

Table 5: Shapiro-Wilk Normality Test

In table 5, as the p value was significant ($p < 0.05$) for each combination of independent and dependent variable, we reject the null hypothesis and conclude that data does not follow univariate normality.

	Test	Statistic	p-value	Result
1	Skewness	66.3379	0	NO
2	Kurtosis	-13.0427	0	NO
3	MV Normality			NO

Table 6: Mardia Normality Test

In table 6, as the p value was significant ($p < 0.05$) for Mardia's Skewness and Kurtosis test, the null hypothesis has been rejected and concluded that data does not follow multivariate normality.

According to the multivariable central limit theorem, if the sample size is large for each combination of independent and dependent variables, then assumptions about multivariable normalization can be assumed.

Homogeneity of the variance-covariance matrices:

Chi-Sq (approx.)	df	p-value
13.914	3	0.003024

Table 7: Box's M-test for Homogeneity of Covariance Matrices

Multivariate Outliers:

Mahalanobis distance test conducted and resulted in no outliers among popularity and energy.

Linearity Assumption:

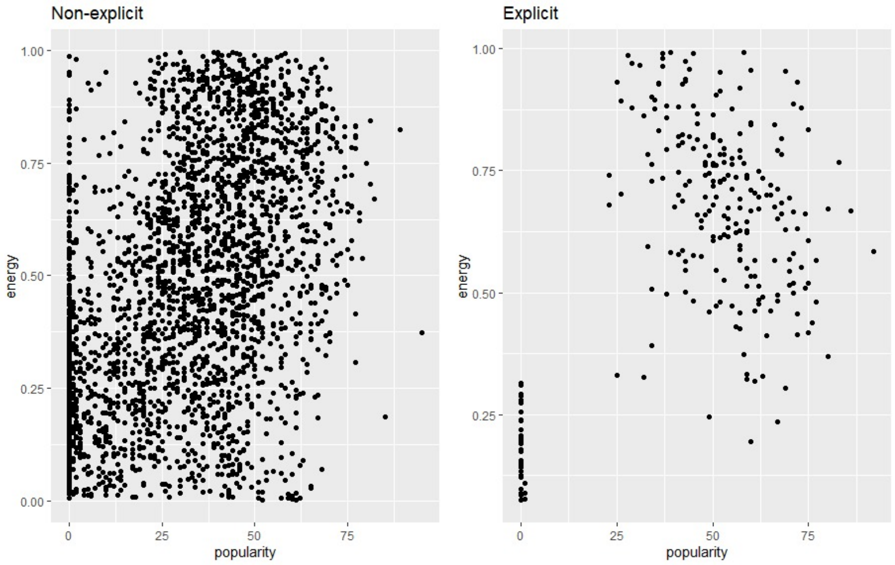


Figure 3.3: Scatter Plot of energy versus popularity for explicit

The Figure 3.3 indicates that dependent variables have a linear relationship for each group in the independent variable.

Multicollinearity Assumption:
Correlation test between the dependent variables gives correlation of 0.48, as the correlation coefficient between the dependent variable was less than 0.9, there is no multicollinearity.

3.3 Principal Component Analysis

	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9	PC10	PC11	PC12	PC13
valence	-0.19	0.53	-0.10	0.32	-0.03	0.32	-0.01	0.03	-0.20	0.38	0.49	-0.01	-0.21
year	-0.39	-0.31	-0.15	-0.21	0.02	-0.12	0.02	0.17	-0.26	-0.06	0.15	-0.68	-0.29
acousticness	0.42	0.06	-0.11	0.10	-0.00	-0.00	0.12	-0.20	-0.30	-0.54	0.45	-0.21	0.34
danceability	-0.22	0.48	-0.26	-0.22	0.02	0.26	-0.15	0.10	-0.35	-0.30	-0.52	-0.02	0.17
duration_ms	-0.00	-0.35	0.28	-0.11	-0.07	0.59	-0.57	-0.31	-0.13	-0.01	0.08	0.00	-0.01
energy	-0.43	-0.00	0.20	0.19	-0.03	0.14	0.06	0.15	0.36	0.02	0.07	-0.26	0.70
instrumentalness	0.26	-0.19	-0.06	0.26	-0.08	0.25	-0.14	0.83	0.03	-0.19	0.01	0.08	-0.10
key	-0.02	0.05	0.03	-0.10	-0.99	-0.07	0.06	-0.02	-0.02	-0.00	-0.01	0.00	-0.00
liveness	-0.01	0.05	0.77	-0.02	0.06	0.05	0.40	0.14	-0.45	-0.00	-0.10	0.03	-0.04
loudness	-0.41	0.03	0.09	0.18	-0.00	0.13	0.17	-0.16	0.31	-0.64	0.08	0.21	-0.40
popularity	-0.38	-0.28	-0.19	-0.21	0.02	-0.13	0.03	0.13	-0.34	0.01	0.32	0.61	0.25
speechiness	0.04	0.39	0.32	-0.53	0.06	-0.25	-0.39	0.23	0.24	-0.14	0.34	0.01	-0.04
tempo	-0.15	0.04	0.16	0.56	-0.03	-0.53	-0.52	-0.05	-0.25	-0.10	-0.12	0.00	0.02

Table 8: Principle Components

	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9	PC10	PC11	PC12	PC13
Standard deviation	1.9821	1.3248	1.0871	1.0384	0.9990	0.9561	0.9346	0.8563	0.7871	0.6001	0.5513	0.3641	0.3480
Proportion of Variance	0.3022	0.1350	0.0909	0.0829	0.0768	0.0703	0.0672	0.0564	0.0477	0.0277	0.0234	0.0102	0.0093
Cumulative Proportion	0.3022	0.4372	0.5281	0.6111	0.6878	0.7581	0.8253	0.8818	0.9294	0.9571	0.9805	0.9907	1.0000

Table 9: Importance of Components

The table 9 shows the standard deviation, the proportion of variance explained by each principal component, and the cumulative proportion of variation explained. For example, it can be observed that the first eight components explain 88.18 percent of the variance in data, which is great. In the figure 3.4, it shows that explained variance of each principle components.

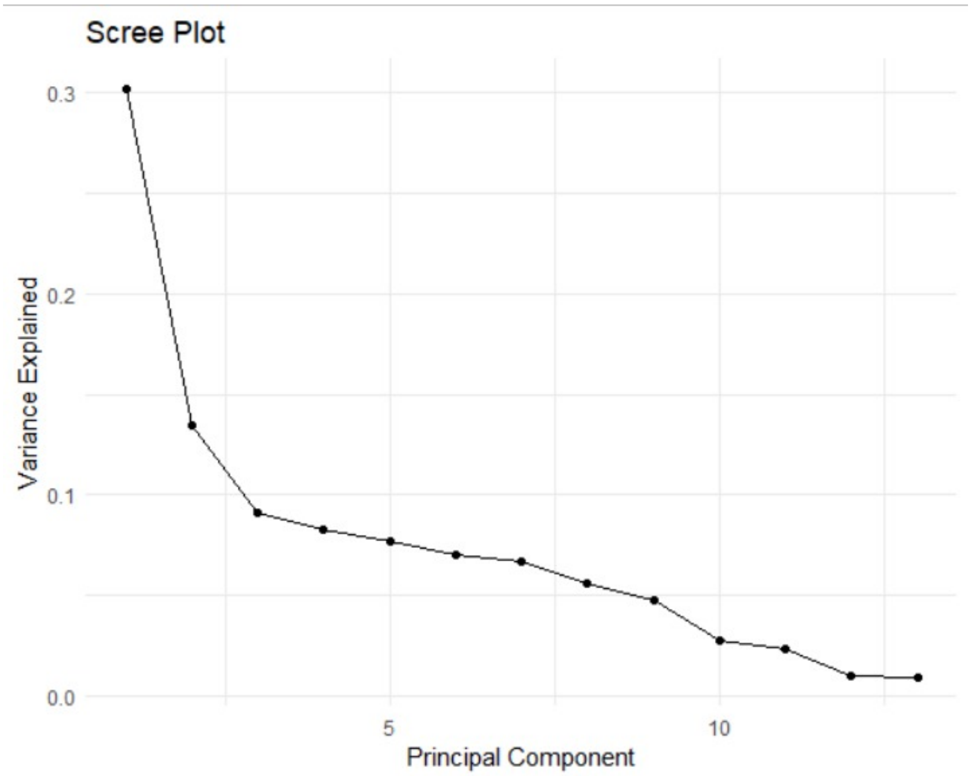


Figure 3.4: Scree Plot

VALIDATION: RMSEP												
Cross-validated using 10 random segments.												
	Intercept	1 comps	2 comps	3 comps	4 comps	5 comps	6 comps	7 comps	8 comps	8 comps	10 comps	11 comps	12 comps
CV	21.83	16.99	15.42	15.18	14.37	14.31	13.89	13.82	13.69	11.69	11.69	10.84	10.84
adjCV	21.83	16.99	15.42	15.18	14.37	14.31	13.89	13.82	13.69	11.69	11.69	10.84	10.84
TRAINING: % variance explained													
X		28.70	42.26	51.88	60.39	68.69	76.13	83.39	89.42	93.76	96.76	98.98	100.00
popularity		39.41	50.12	51.64	56.65	57.01	59.50	59.92	60.69	71.33	71.34	75.32	75.34

Table 10: Principal Component Regression Model

From Table 10, if only the intercept term is used in the model, the test RMSE is 21.83. If the first principal component added, the test RMSE drops to 16.99. If the second principal component added, the test RMSE drops to 15.42. It can be seen that adding additional principal components actually leads to an increase in test RMSE.

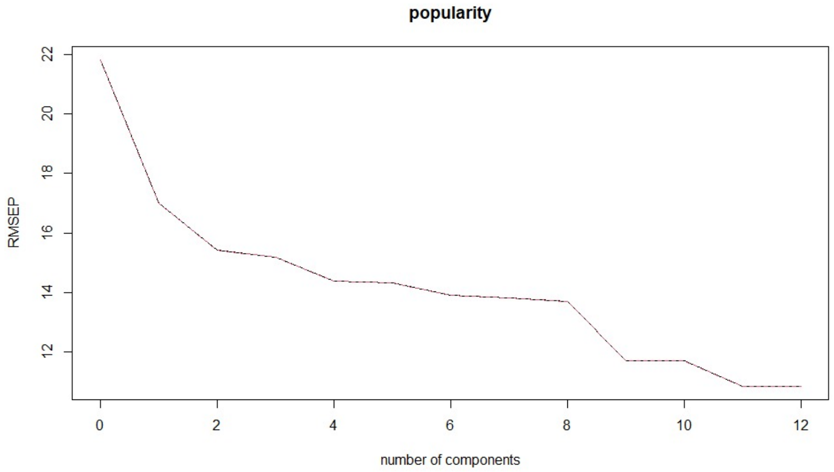


Figure 3.5: RMSEP Plot of the Model

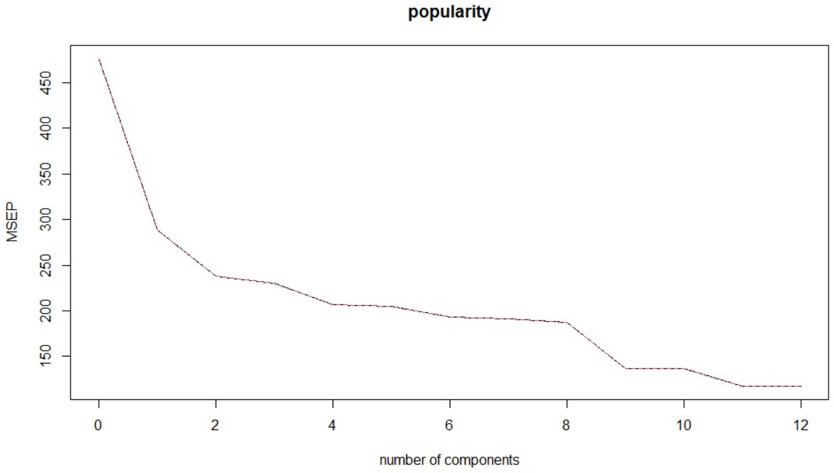


Figure 3.6: MSEP Plot of the Model

In Figures 3.5 and 3.6, adding more components improves the model fit for all the components.

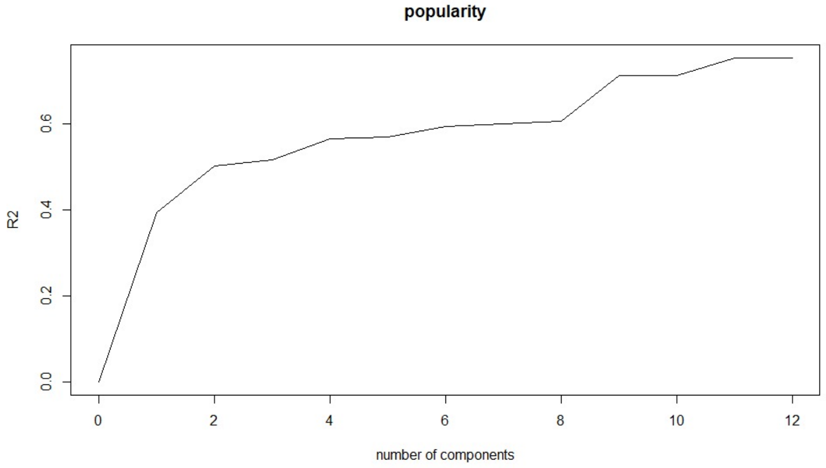


Figure 3.7: R^2 *Plot of the Model*

According the Figure 3.7, explained variance increases with each added component, in other words, model fit improves.

Then, the data set divided into train and test groups to test predictions of the conducted model on the test group with respect to 8 principal components.

Test for RMSE turns out to be 13.79935. This is the average deviation between the predicted value for popularity and the observed value for popularity for the observations in the testing set.

3.4 Factor Analysis

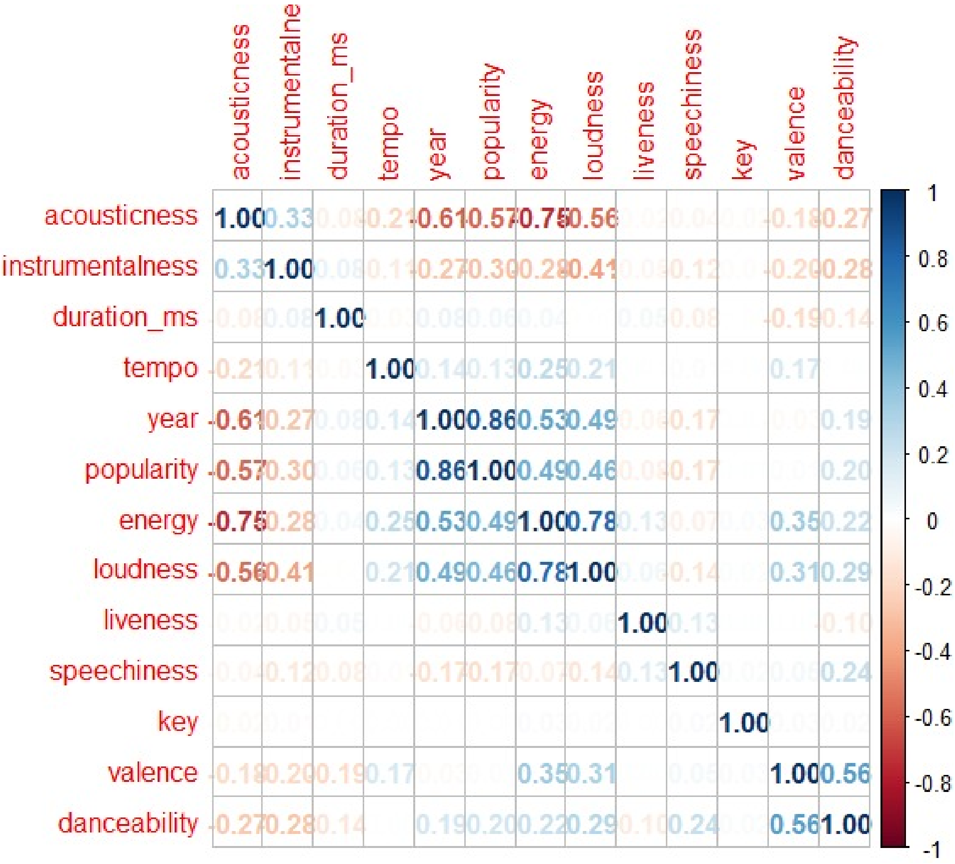


Figure 3.8: Correlation Plot of Numeric Variables

Kaiser-Meyer-Olkin factor adequacy

Overall MSA = 0.7

MSA for each item:

valence	year	acousticness	danceability	duration_ms	energy	instrumentalness	key	liveness	loudness	popularity	speechiness	tempo
0.54	0.73	0.80	0.56	0.71	0.68	0.76	0.79	0.42	0.75	0.74	0.42	0.78

Table 11: Kaiser-Meyer-Olkin Factor Adequacy

Since MSA is greater than 0.5, Factor Analysis on this data can be run. Besides, Bartletts test of sphericity should be significant.

Chisq	885185.6
df	78
p-value	0

Table 12: Bartlett's Test

The Kaiser-Meyer Olkin (KMO) and Bartletts Test measure of sampling adequacy were used to examine the appropriateness of Factor Analysis. The approximate of Chi-square is

885185.6 with 78 degrees of freedom, which is significant at 0.05 Level of significance. The KMO statistic of 0.7 is also large (greater than 0.50). Hence Factor Analysis is considered as an appropriate technique for further analysis of the data.

Then, number of factors were needed to be decided and there are several way to do it. It can be done with using visual ways or formal ways. The scree plot, which graphs the Eigenvalue against each factor, is used to determine number of factors visually.

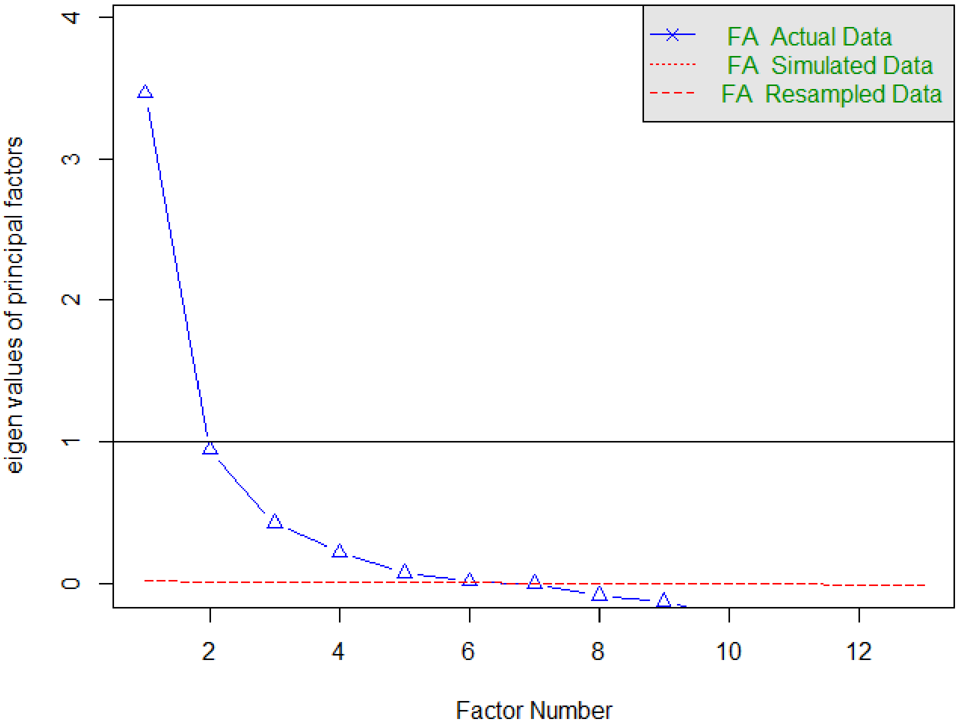


Figure 3.9: Parallel Analysis Scree Plot

And the parallel analysis suggests that the number of factors as 6.

Uniquenesses:						
valence	year	acousticness	danceability	duration_ms	energy	instrumentalness
0.196	0.121	0.307	0.005	0.923	0.005	0.657
key	liveness	loudness	popularity	speechiness	tempo	
0.998	0.904	0.005	0.144	0.542	0.884	

Loadings:						
	Factor1	Factor2	Factor3	Factor4	Factor5	Factor6
valence		0.321	0.800		0.232	
year	0.884	0.204	-0.110	0.154		-0.111
acousticness	-0.547	-0.581		-0.154		-0.165
danceability	0.172		0.429	0.142	0.835	0.236
duration_ms			-0.256			
energy	0.356	0.893		0.251		
instrumentalness	-0.289		-0.213	-0.388		-0.239
key						
liveness		0.174			-0.166	0.170
loudness	0.281	0.532	0.109	0.768		-0.157
popularity	0.892	0.147		0.139		-0.122
speechiness	-0.100					0.660
tempo	0.131	0.192	0.178		-0.153	
SS loadings	2.235	1.673	1.004	0.903	0.831	0.663
Proportion Var	0.172	0.129	0.077	0.069	0.064	0.051
Cumulative Var	0.172	0.301	0.378	0.447	0.511	0.562

Table 13: Varimax Solution

Varimax solution or rotation enables to interpret the factor loadings. For example, the first factor is dominated by the year and popularity. Second factor reflects energy. Dimensionality can be effectively reduced from 13 to 6 while only losing about 44% of the variance. Factor 1 accounts for 17.2% of the variance; Factor 2 accounts for 12.9% of the variance and it continues like this.

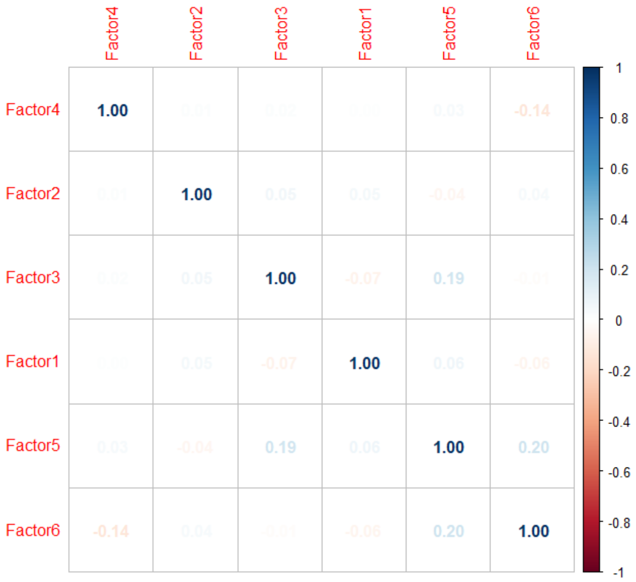


Figure 3.10: Parallel Analysis Scree Plot

As it can be seen, they are almost uncorrelated which guarantees that no multicollinearity problem in linear regression after the deciding the factor number as 6.

3.5 Canonical Correlation Analysis

For canonical analysis, variables were divided to characteristic and structure groups. Characteristic variables are valued between 0 and 1 values, and structure group consist of variables such as publication year of the song, duration and loudness etc.

Following two graphs show the relationships of the variables for both characteristic and structure group.

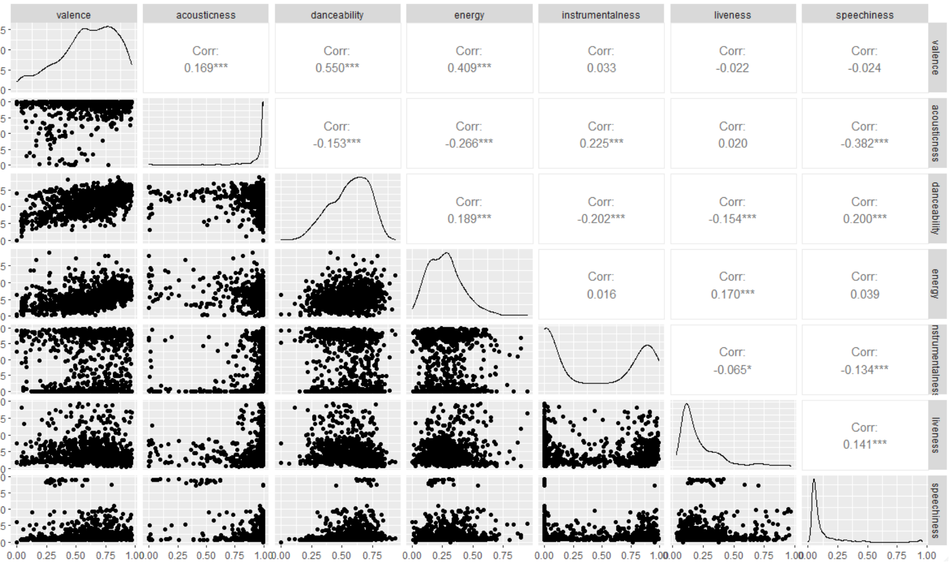


Figure 3.11: Correlation Plot of Characteristics Group

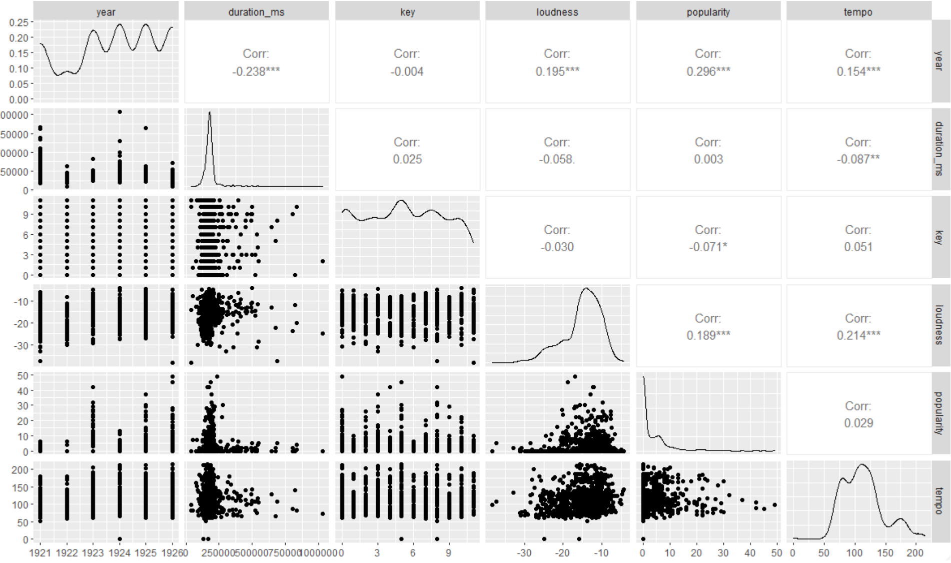


Figure 3.12: Correlation Plot of Structure Group

Following graph, Figure 3.13, shows the correlation between each group and between groups.

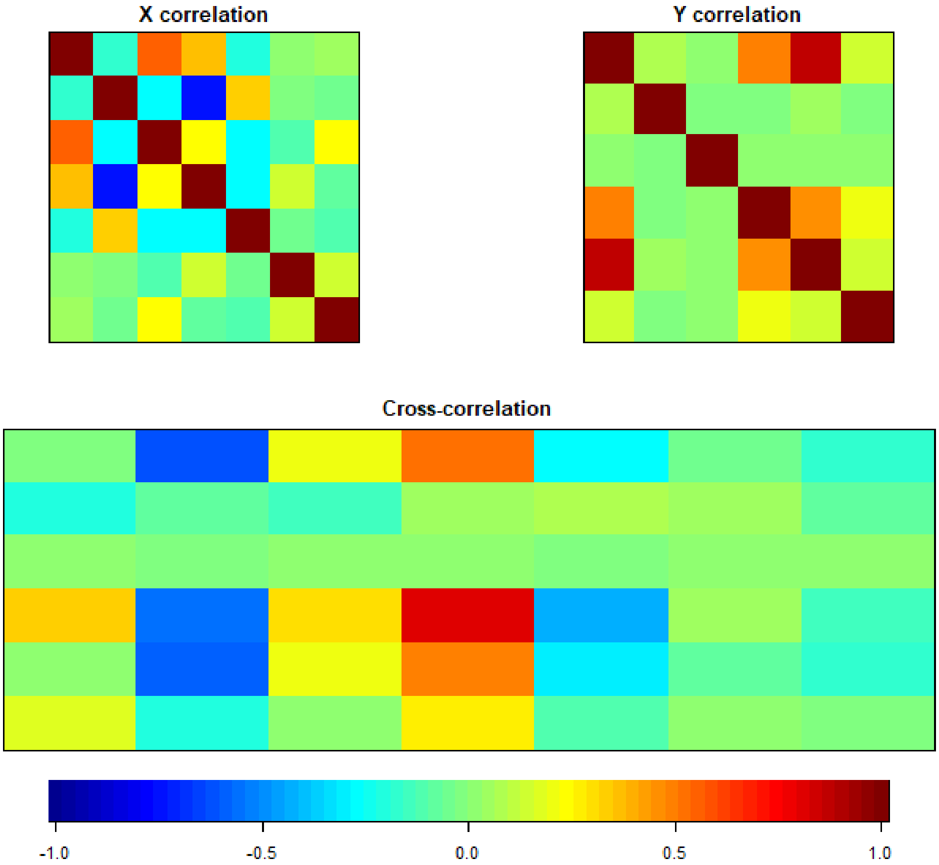


Figure 3.13: Correlation Between Each Group and Between Groups

X represents the characteristic group and Y represents the structure group. Then, canonical correlations were calculated for the groups.

Cancor Approach	0.855181579	0.514760470	0.223813013	0.205682726	0.076640535	0.004254229
CC Approach	0.855181579	0.514760470	0.223813013	0.205682726	0.076640535	0.004254229

Table 14: Canonical Correlations for Different Approaches

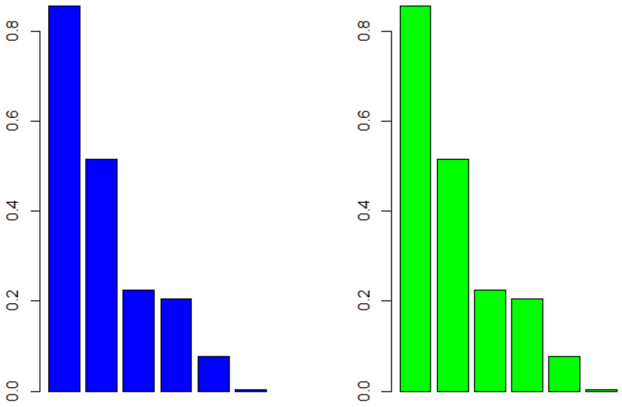


Figure 3.14: Canonical Correlations for cancor() and cc()

As it can be seen, the canonical correlations are the same for both approaches.

In addition, some tests for canonical dimensions were conducted which are Wilks, Hotelling and Pillai tests.

	stat	approx	df1	df2	p.value
1 to 6:	0.178	8457.214	42	800376.0	0
2 to 6:	0.664	2445.611	30	682566.0	0
3 to 6:	0.904	870.829	20	565956.4	0
4 to 6:	0.952	705.248	12	451479.2	0
5 to 6:	0.994	168.309	6	341288.0	0
6 to 6:	0.999	1.544	2	170645.0	0.213

Table 15: Wilks' Lambda, using F-approximation (Rao's F)

	stat	approx	df1	df2	p.value
1 to 6:	3.185	12941.91	42	1023830	0
2 to 6:	0.463	2635.47	30	1023842	0
3 to 6:	0.102	877.39	20	1023854	0
4 to 6:	0.050	712.45	12	1023866	0
5 to 6:	0.006	168.55	6	1023878	0
6 to 6:	0.000	1.54	2	1023890	0.213

Table 16: Hotelling-Lawley Trace, using F-approximation

	stat	approx	df1	df2	p.value
1 to 6:	1.094	5439.74	42	1023870	0
2 to 6:	0.363	2199.52	30	1023882	0
3 to 6:	0.098	852.61	20	1023894	0
4 to 6:	0.048	690.95	12	1023906	0
5 to 6:	0.006	167.74	6	1023918	0
6 to 6:	0.000	1.54	2	1023930	0.213

Table 17: Pillai-Bartlett Trace, using F-approximation

According to all of these tests, dimensions 1 to 5 are significant while dimension 6 is not, with respect to p-values.

Then, standardized canonical coefficients were calculated.

	1	2	3	4	5	6
1	-0.19	0.71	-0.49	-0.85	-0.11	0.19
2	-0.05	1.01	0.51	0.96	-0.15	0.14
3	0.22	-0.28	-0.25	1.12	0.32	0.21
4	0.84	0.70	0.90	0.53	0.25	0.26
5	-0.24	-0.16	0.45	-0.13	0.52	0.67
6	-0.04	0.04	0.34	0.15	-0.15	-0.51
7	-0.21	0.28	0.07	-0.21	0.83	-0.39

Table 18: Standardized Canonical Coefficients of Characteristics Group

	1	2	3	4	5	6
1	0.26	-0.86	0.67	0.09	1.60	0.53
2	0.02	-0.26	0.80	-0.17	-0.48	-0.22
3	0.01	0.03	-0.04	-0.03	0.35	-0.94
4	0.77	0.75	0.27	0.33	-0.14	-0.02
5	0.11	-0.17	-1.13	-0.17	-1.53	-0.51
6	0.04	0.18	-0.03	-1.00	0.11	0.08

Table 19: Standardized Canonical Coefficients of Structure Group

The standardized canonical coefficients were interpreted similarly to the standardized regression coefficients. When the other variables in the model was maintained constant, a one standard deviation increase in year refers to a 0.26 standard deviation increase in the score on the first canonical variate for set 2.

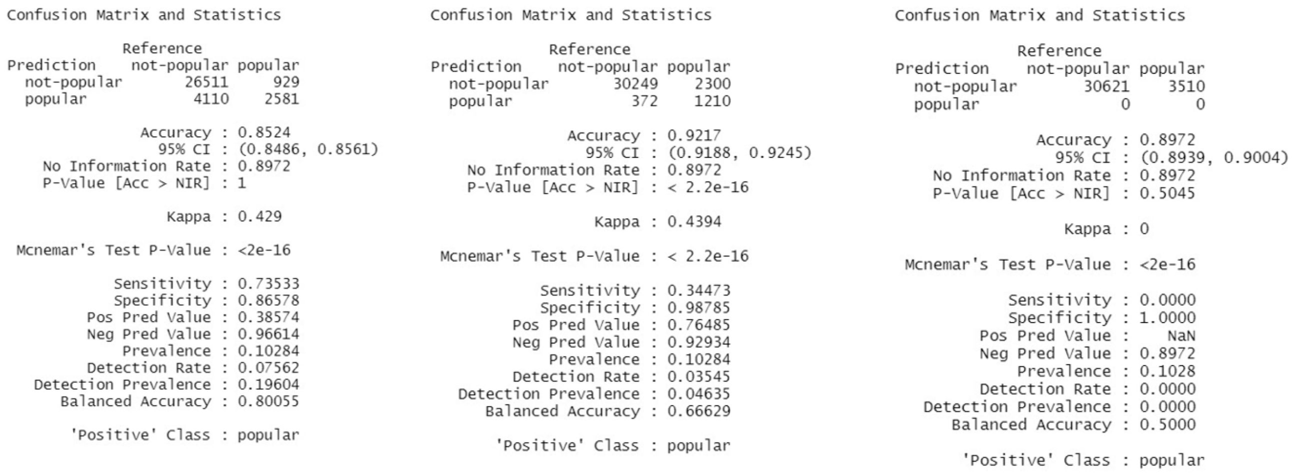
3.6 Logistic Regression

The dataset's popularity column was separated into two groups: songs with a popularity score of 60 or higher were deemed popular, while songs with a score of 60 or lower were considered unpopular. Popularity was chosen as a response. To do logistic regression, 80 percent of the data was utilized for training and the remaining 20 percent was used for testing. The training set consists of 136522 observations with 15 variables, whereas the test set consists of 34131 observations with 15 variables.

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-181.0293	1.9154	-94.51	0.0000
valence	-0.2757	0.0550	-5.01	0.0000
year	0.0901	0.0009	95.30	0.0000
acousticness	-0.0932	0.0513	-1.82	0.0693
danceability	0.9234	0.0811	11.38	0.0000
duration_ms	-0.0000	0.0000	-11.58	0.0000
energy	-0.5115	0.0891	-5.74	0.0000
explicit1	0.3542	0.0319	11.11	0.0000
instrumentalness	-0.7842	0.0595	-13.19	0.0000
key	-0.0030	0.0030	-0.99	0.3200
liveness	-0.4553	0.0703	-6.47	0.0000
loudness	0.0213	0.0042	5.08	0.0000
mode1	-0.1189	0.0228	-5.21	0.0000
speechiness	-1.1000	0.1261	-8.72	0.0000
tempo	0.0004	0.0004	1.21	0.2256

Table 20: Logistic Regression Model

All the main effects were significant at 5% significance, except acousticness, key and tempo. Acousticness was significant at 10% significance. Three different cutoffs were made to asses the classification results which are 0.2, 0.5 and 0.8.



```
Confusion Matrix and Statistics

                Reference
Prediction    not-popular popular
  not-popular       26511     929
  popular            4110    2581

               Accuracy : 0.8524
                 95% CI : (0.8486, 0.8561)
    No Information Rate : 0.8972
    P-Value [Acc > NIR] : 1

                  Kappa : 0.429

 Mcnemar's Test P-Value : <2e-16

            Sensitivity : 0.73533
            Specificity : 0.86578
         Pos Pred Value : 0.38574
         Neg Pred Value : 0.96614
             Prevalence : 0.10284
         Detection Rate : 0.07562
   Detection Prevalence : 0.19604
      Balanced Accuracy : 0.80055

       'Positive' Class : popular
```

```
Confusion Matrix and Statistics

                Reference
Prediction    not-popular popular
  not-popular       30249    2300
  popular             372    1210

               Accuracy : 0.9217
                 95% CI : (0.9188, 0.9245)
    No Information Rate : 0.8972
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.4394

 Mcnemar's Test P-Value : < 2.2e-16

            Sensitivity : 0.34473
            Specificity : 0.98785
         Pos Pred Value : 0.76485
         Neg Pred Value : 0.92934
             Prevalence : 0.10284
         Detection Rate : 0.03545
   Detection Prevalence : 0.04635
      Balanced Accuracy : 0.66629

       'Positive' Class : popular
```

```
Confusion Matrix and Statistics

                Reference
Prediction    not-popular popular
  not-popular       30621    3510
  popular               0       0

               Accuracy : 0.8972
                 95% CI : (0.8939, 0.9004)
    No Information Rate : 0.8972
    P-Value [Acc > NIR] : 0.5045

                  Kappa : 0

 Mcnemar's Test P-Value : <2e-16

            Sensitivity : 0.0000
            Specificity : 1.0000
         Pos Pred Value :    NaN
         Neg Pred Value : 0.8972
             Prevalence : 0.1028
         Detection Rate : 0.0000
   Detection Prevalence : 0.0000
      Balanced Accuracy : 0.5000

       'Positive' Class : popular
```

Figure 3.15: Cutoff Results for 0.2, 0.5, 0.8 respectively

For the cutoff 0.2, positive predictive value is too low, so it cannot classify popular songs correctly, but it is better in predicting non-popular songs. It does not seem to perform well. For cutoff 0.5, although it classifies most of the songs correctly, its sensitivity is too low. This cutoff does not perform well either. And finally, for cutoff 0.8 there is no popular songs in the test dataset, so it performs the worst. Logistic regression model did not perform well enough to classify popular songs in all of the cutoffs.

3.7 Cluster Analysis

First, before clustering, numeric variables were selected and data was scaled for standard normality. K-means clustering was done for songs that have popularity level higher than 70, as stated in data description popularity level ranges from 0 to 100.

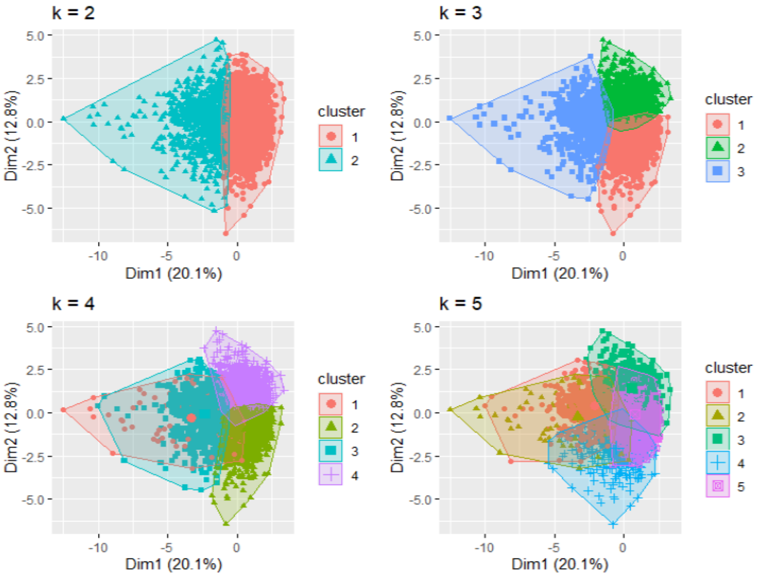


Figure 3.16: Clustering Method

In Figure 3.16, results of k-means clustering were illustrated for 4 different number of clusters, which are 2, 3, 4 and 5, respectively. Then, different methods were tried to find optimal number of clusters for the given observations.

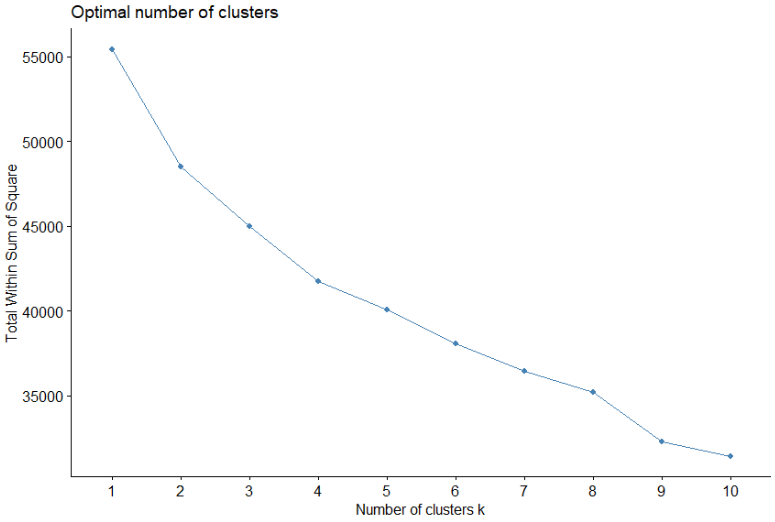


Figure 3.17: Wss Method

Figure 3.17 suggests that as the cluster number increases total sum of squares decreases steadily so optimal number of clusters might be greater than ten clusters.

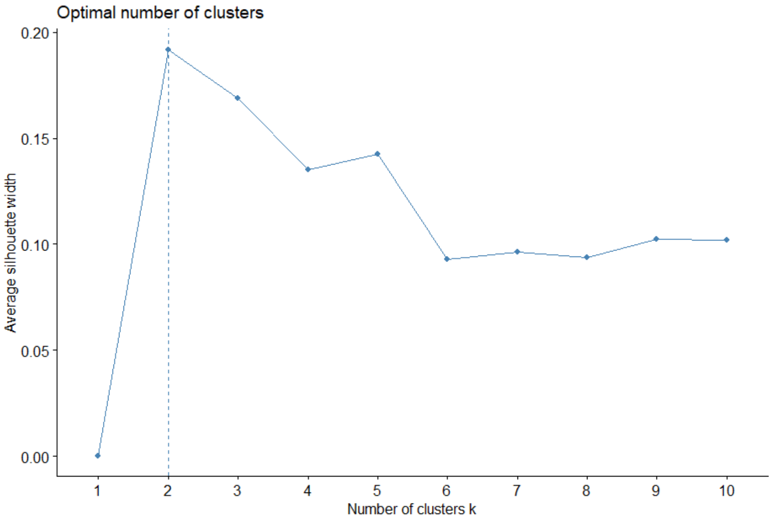


Figure 3.18: Silhouette Method

In Figure 3.18, average silhouette width method for selecting clusters suggest two cluster as an optimal number of clusters.

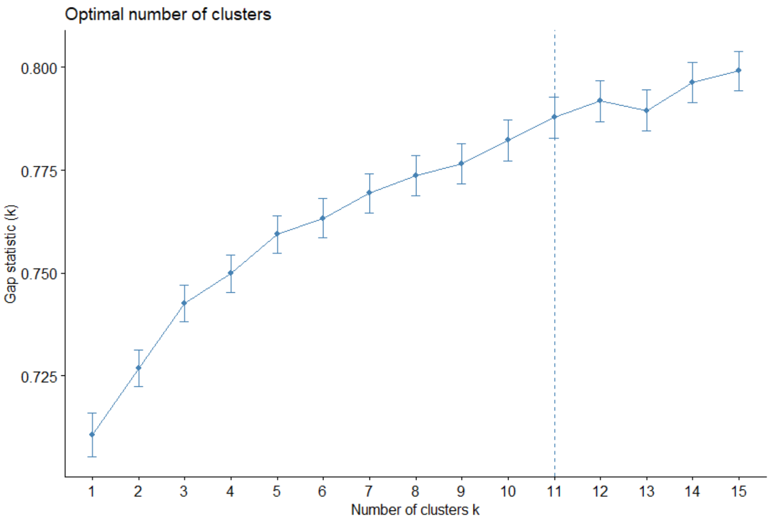


Figure 3.19: Gap Stat Method

In Figure 3.19, gap statistics method for selecting optimal number of clusters suggests that optimal number should be eleven. Hence, eleven numbers of clusters were chosen in the final K-means clustering as illustrated below in Figure 3.20.

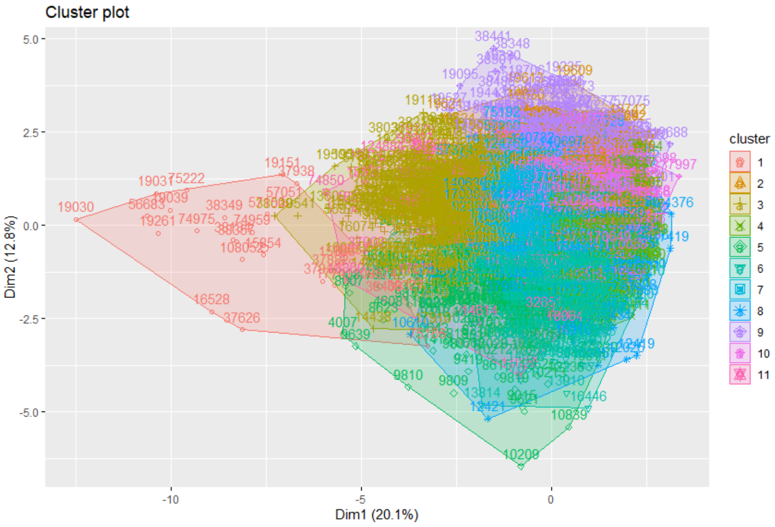


Figure 3.20: Cluster Plot when K=11

3.8 Decision Tree

For conducting decision tree, 10,000 observations were sampled from the data because decision tree algorithm could not finish computing due to insufficient memory for all the 170,653 observations. Popularity was the dependent variable and was divided into two groups with cutting value of 50, to put it simpler, popularity higher than 50 was considered popular and less than 50 was considered as not popular. Sampled data was divided into training and test sets with 80 percent being the training set. For both sets, proportions were 0.83 and 0.17 for non-popular and popular songs, respectively.

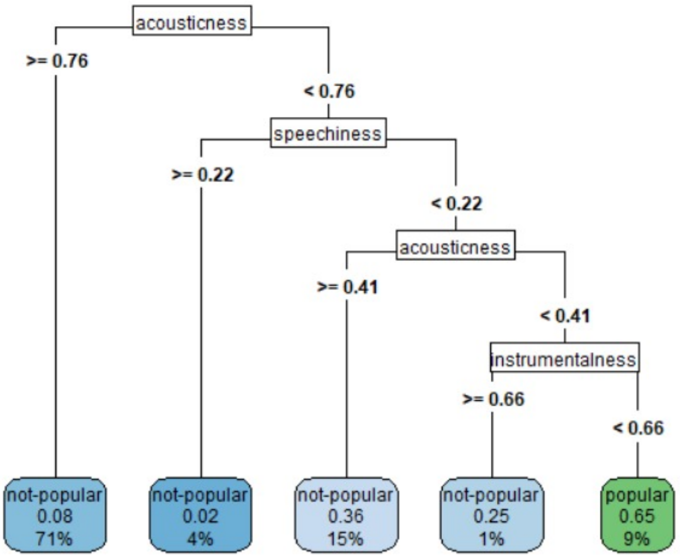


Figure 3.21: Decision Tree for Popularity

According to decision tree shown in Figure 3.21, to be a popular song it needs to have less acousticness (vocals), less speechness and less instrumentalness. Classified popular songs were only 9%, so it seems to classify only a small proportion of popular songs.

	not-popular	popular
not-popular	1606	65
popular	223	106

Table 21: Confusion Matrix for Decision Tree

Confusion matrix for test data as shown in Table 21, indicates classification of non-popular songs were mostly correct but it failed to classify most of the popular songs. Accuracy for the decision tree was 85%, sensitivity was 62% and specificity was 87% but positive predictive value was only 32%. Classification was consistent for both train and test datasets.

3.9 Random Forest

To conduct random forest algorithm, data was split into train and test datasets with train dataset consisting 70% of the data. Algorithm could not run on whole train dataset due to inability to allocate enough storage space, so to proceed with the algorithm 2,000 observations were randomly sampled from the dataset.

	not-popular	popular	class.error
not-popular	1595	68	0.04
popular	131	206	0.39

Table 22: Confusion Matrix for Random Forest

The number of trees is 500 in the model and number of variables tried at each split is 3. For Table 22, classification error in non-popular songs is 0.0408 i.e., 4.08%, in popular songs is 0.389 i.e., 38.87%. Random forest algorithm classification was better in classifying popular and non-popular songs compared to logistic regression and decision tree algorithm. Accuracy of random forest for test dataset was 89%, sensitivity was 0.96 and specificity was 60%, which was a few less than the values for train dataset with train set accuracy and specificity being 93% and 75%, respectively and all others were very close.

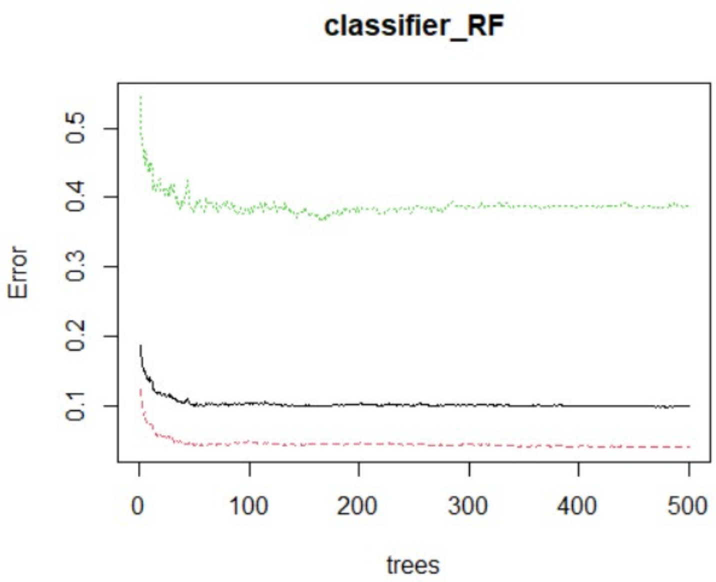


Figure 3.22: Error Rates for Random Forest

Figure 3.22 illustrates that for both non-popular and popular as well as overall errors were stabilized with an increase in the number of trees. Error rate has stayed high for popular songs.

	MeanDecreaseGini
valence	27.47
year	198.71
acousticness	64.31
danceability	27.48
duration_ms	30.36
energy	41.36
explicit	0.13
instrumentalness	25.92
key	15.86
liveness	25.60
loudness	32.46
mode	2.84
speechiness	38.95
tempo	27.17

Table 23: Importance of Factors in Random Forest

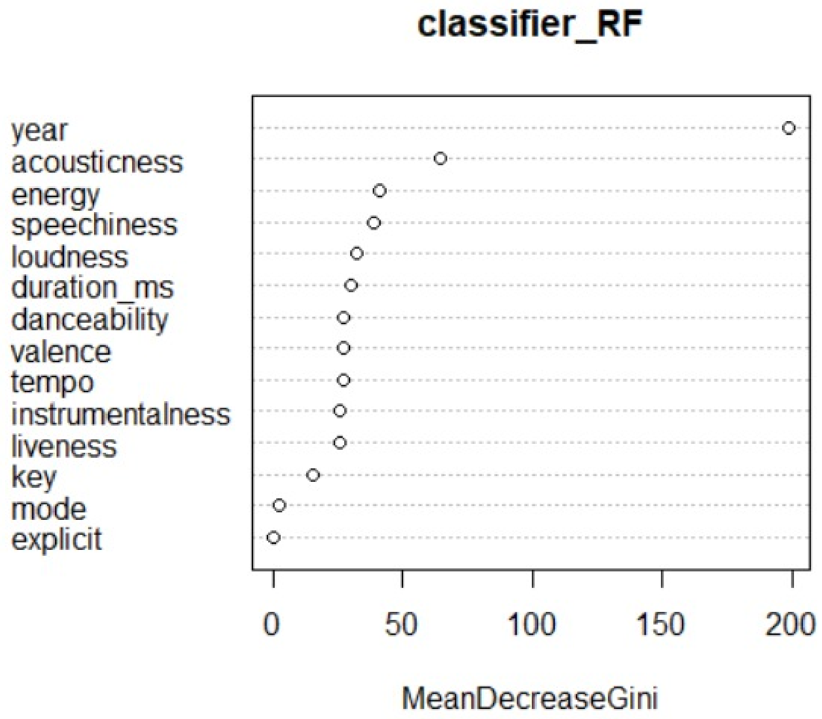


Figure 3.23: Importance of Factors in Random Forest

According to Random Forest algorithm, as shown in the Table 23 and Figure 3.23, the most important factor in songs' popularity is year of the song followed by acousticness, energy and speechiness. Explicit content was the least important factor followed by the mode of the song, both of which being binary variables.

4 Conclusion

To conclude the findings, firstly, this research could not classify popular songs with high accuracy with only the random forest algorithm being able to classify popular songs more correctly than incorrectly. It looks like predicting a popular song by its characteristics and structural features is not easy. However, it is meaningful to conclude that the release year of a song seems to impact the popularity of the song the most, more recent songs being more popular, followed by the negative relationship with vocals (acoustics). It was seen at first that explicit songs were more popular, but Decision Tree and Random Forest algorithms suggest that it is, in fact, not an important factor for popularity when other factors are in the model. Furthermore, instrumentals and too much speech seem to be negatively associated with the popularity of the song. And finally, dividing the popular songs into eleven clusters seems to have interesting results. Clusters mainly consisted of songs that were around the same years and from the same artists and the most interesting observation was that almost all the Christmas songs were in the same cluster.

References

[1] I. Eady and J. D. Wilson, “The influence of music on core learning.” 2004. [Online]. Available: https://link.gale.com/apps/doc/A127013750/AONE?u=anon~df5d0ca5&sid=googleScholar&xid=f31b0bcd

[2] “Spotify dataset 1921-2020,” https://data.world/babarory/spotify-dataset-1921-2020, accessed: 2022-02-10.