TOP HITS ON SPOTIFY FROM 2000 TO 2019

AIM OF THE STUDY:

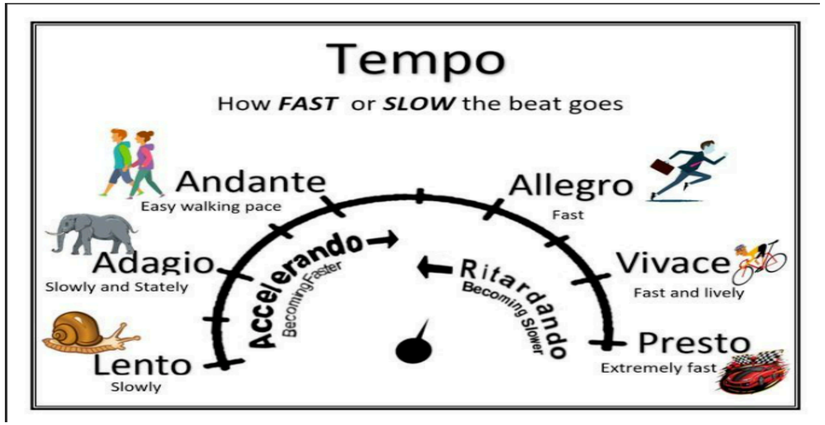
The aim of the study is to analyze the audio features of the top 2000 Spotify tracks released between 2000 and 2019 using multivariate statistical methods. By applying exploratory data analysis and advanced multivariate techniques we investigate relationship, structures and patterns in the dataset.

Presentation Outline

- Data Explanation
- Exploratory Data Analysis
- Inferences About A Mean Vector
- Comparisons of Several Multivariate Means
- Principal Components Analysis ad Principal Componenets Regression
- Factor Analysis and Factor Rotation
- Discrimination and Classification
- Clustering
- Canonical Correlation Analysis

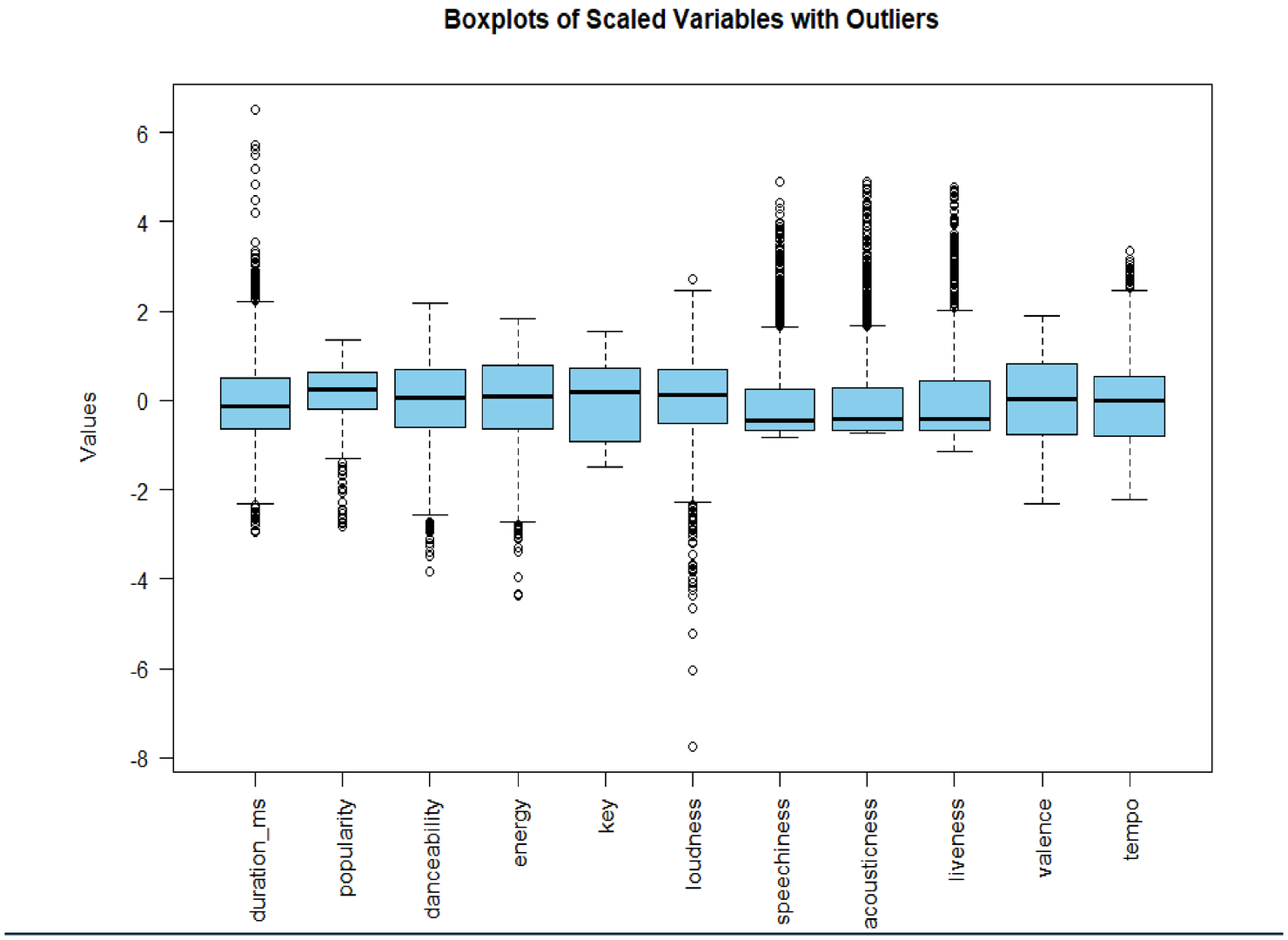
COLUMNS DESCRIPTION

VARIABLE	DESCRIPTION
Popularity	A higher value means the track is more popular
Loudness	Avarage loudness in decibels (-60 to 0 dB)
Acousticness	Confidence level of the track being acoustic (0.0-1.0
Speechiness	Detects spoken word presence (higher values indicate more speech-like content)
Liveness	Detects the presence of an audience in the recording. A value above 0.8 provides strong likelihood that the track is live.
Explicit	Indicates if the content is offensive or unsuitabel
Instrumentalness	Likelihood the track has no vocals (0.0–1.0)
Danceability	Measures how suitable a track for dancing (0.0-1.0)
Key	Musical key of the track
Valance	Positiveness of the track's mood (0.0-1.0)
Energy	Represents a perceptual measure of intensity and activity
Mode	Scale type (1 = Major, 0 = Minor)
Tempo	The overall estimated tempo of a track in beats per minute (BPM)
Duration_ms	Track length in milliseconds (ms)

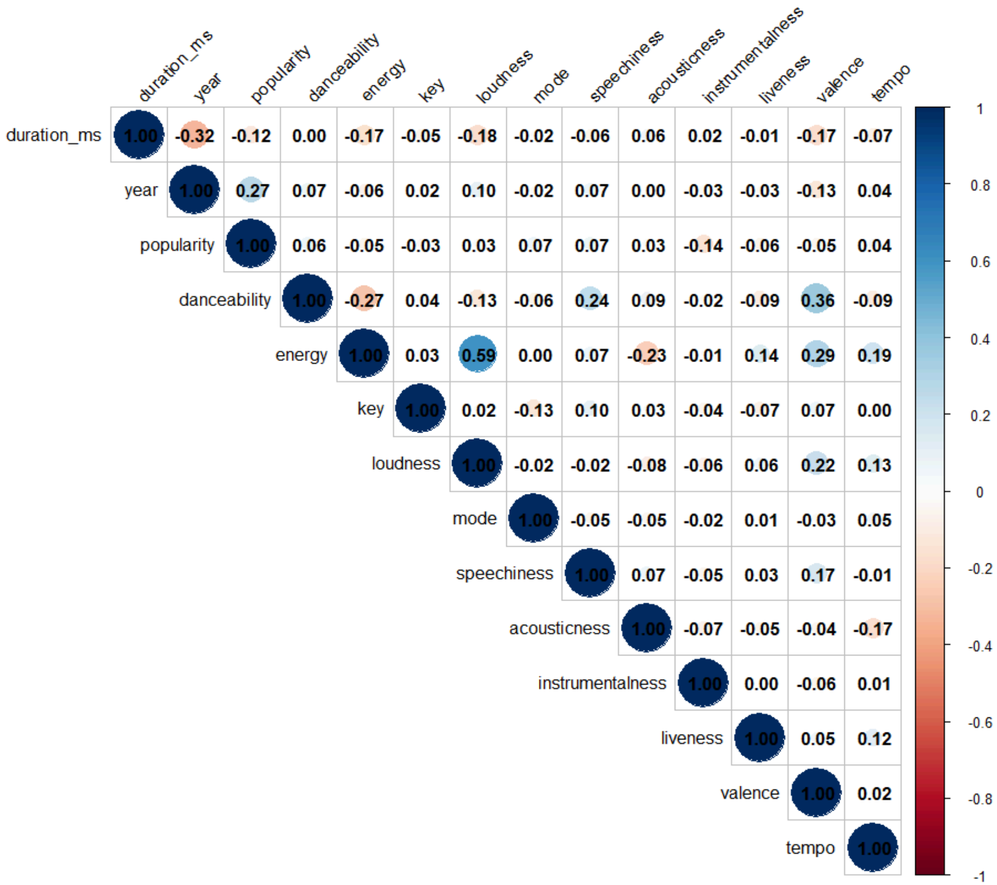


DATA SET INFORMATION

This dataset is from Kaggle and includes audio statistics for the top 2000 Spotify tracks released between 2000-2019. It consist of 18 columns, each describing the track and characteristics of the tracks.



Each box shows that their IQRs and their means. Also whiskers shows range of data within 1.5 times the IQR from Q1 and Q3. Loudness, accoustichness and speechiness have significant outliers. Key and valence don't have any outliers. Key is discrete value. Moreover it has a small range. For valence has max value as 1 min value as 0 again small range. Before making any test outliers are detected and excluded from the data. (For PCA full data was used.)



Correlation Matrix

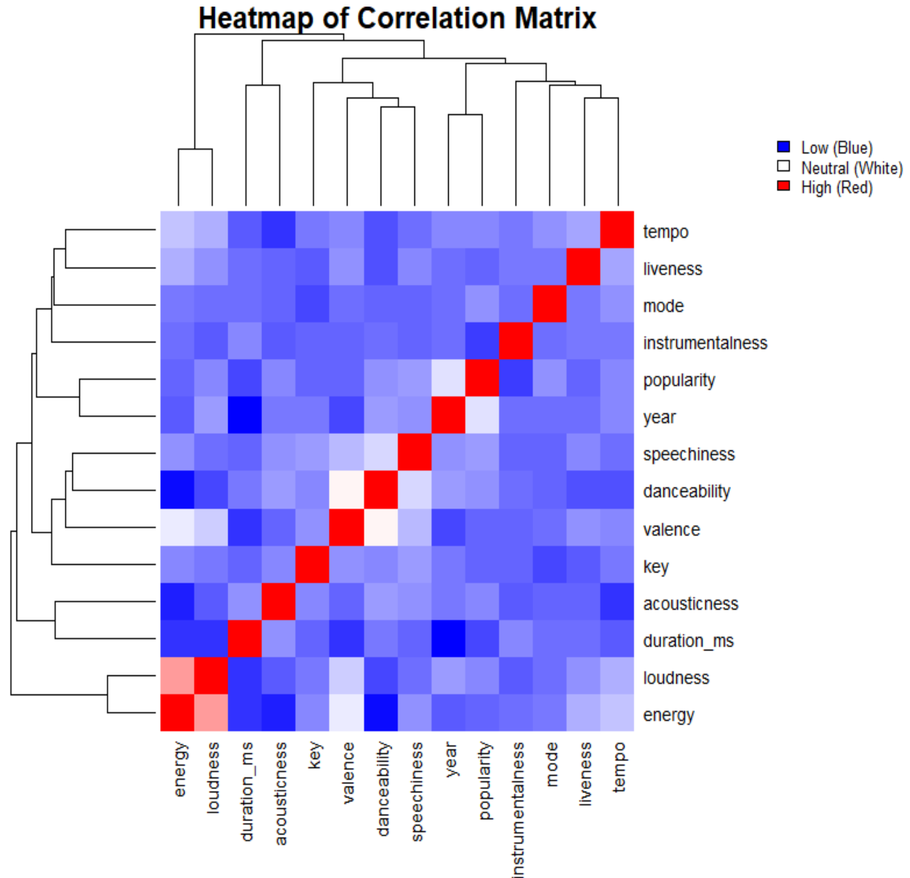
No variables have high correlation coefficient (generally below 0.8). Just energy and loudness has correlation coefficient as 0.59. It is also smaller than 0.8 so there is no problem. So, variables do not have strong linear dependencies on each other. Also some coefficients are positive some of them are negative. Like -0.23 it means that energy and acousticness has weak negative relationship. Also, valence and danceability has coefficient as 0.36. It indicates weak positive relationship.

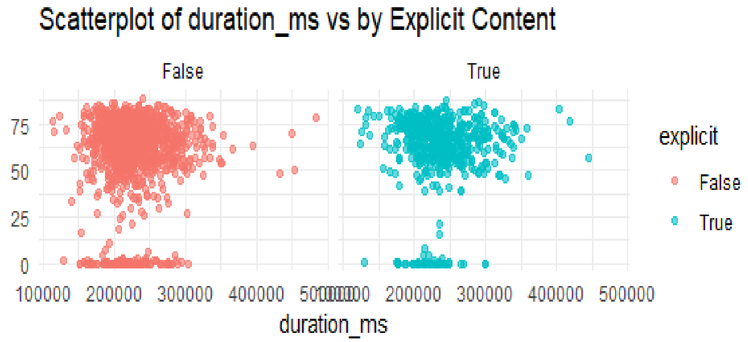
Heatmap

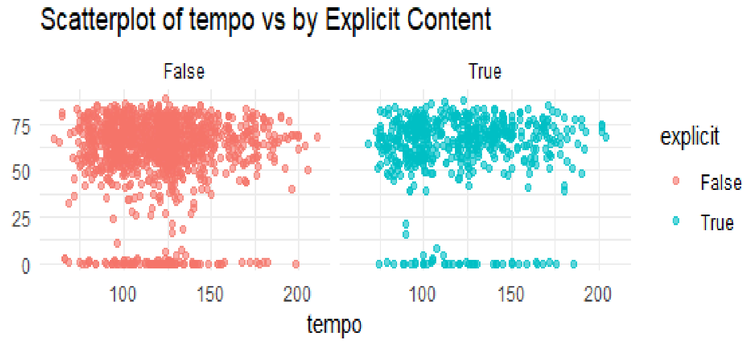
Red boxes are perfect correlations since each variable is perfectly correlated with itself.

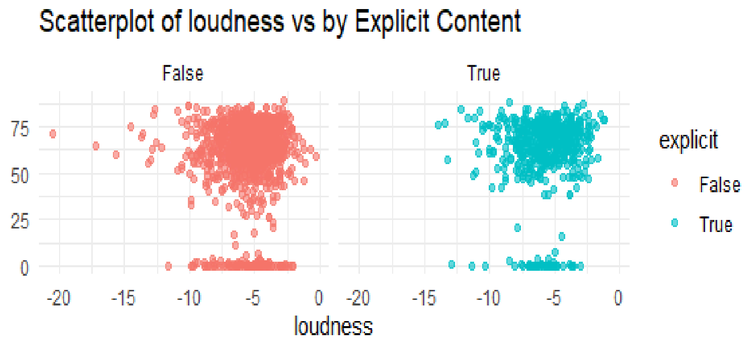
Blue boxes show weak realitionship since as we can see from the correlation matrix our data has mainly uncorrelated variables. eg. popularity and duration

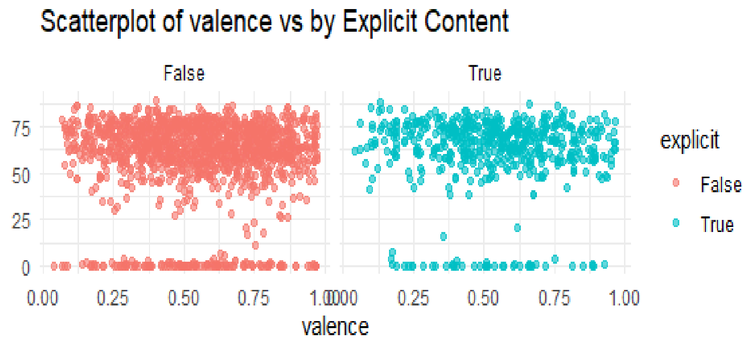
White boxes show no correlation or very low correlation. eg. key and tempo

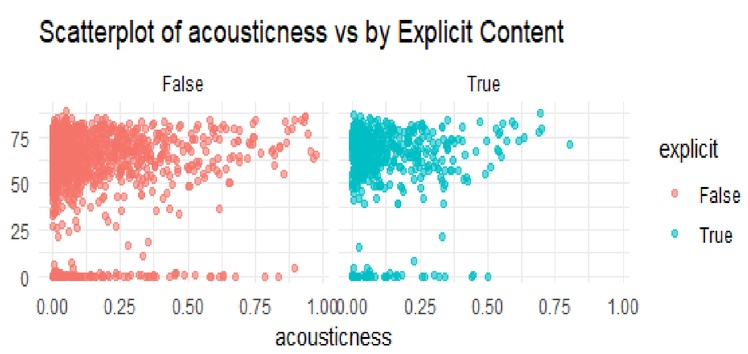


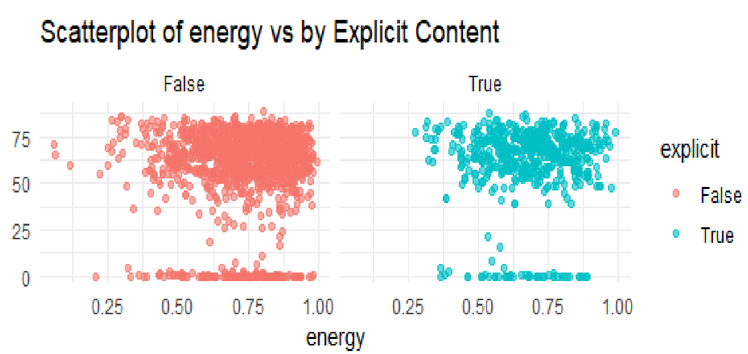


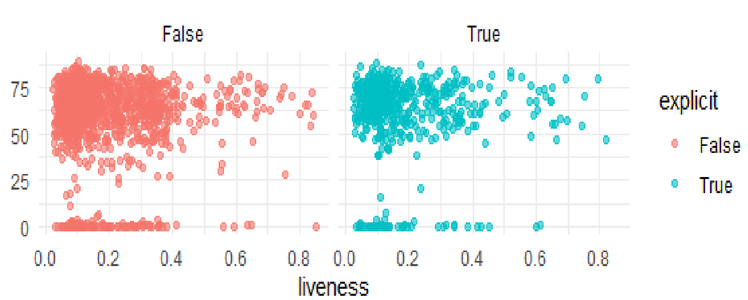










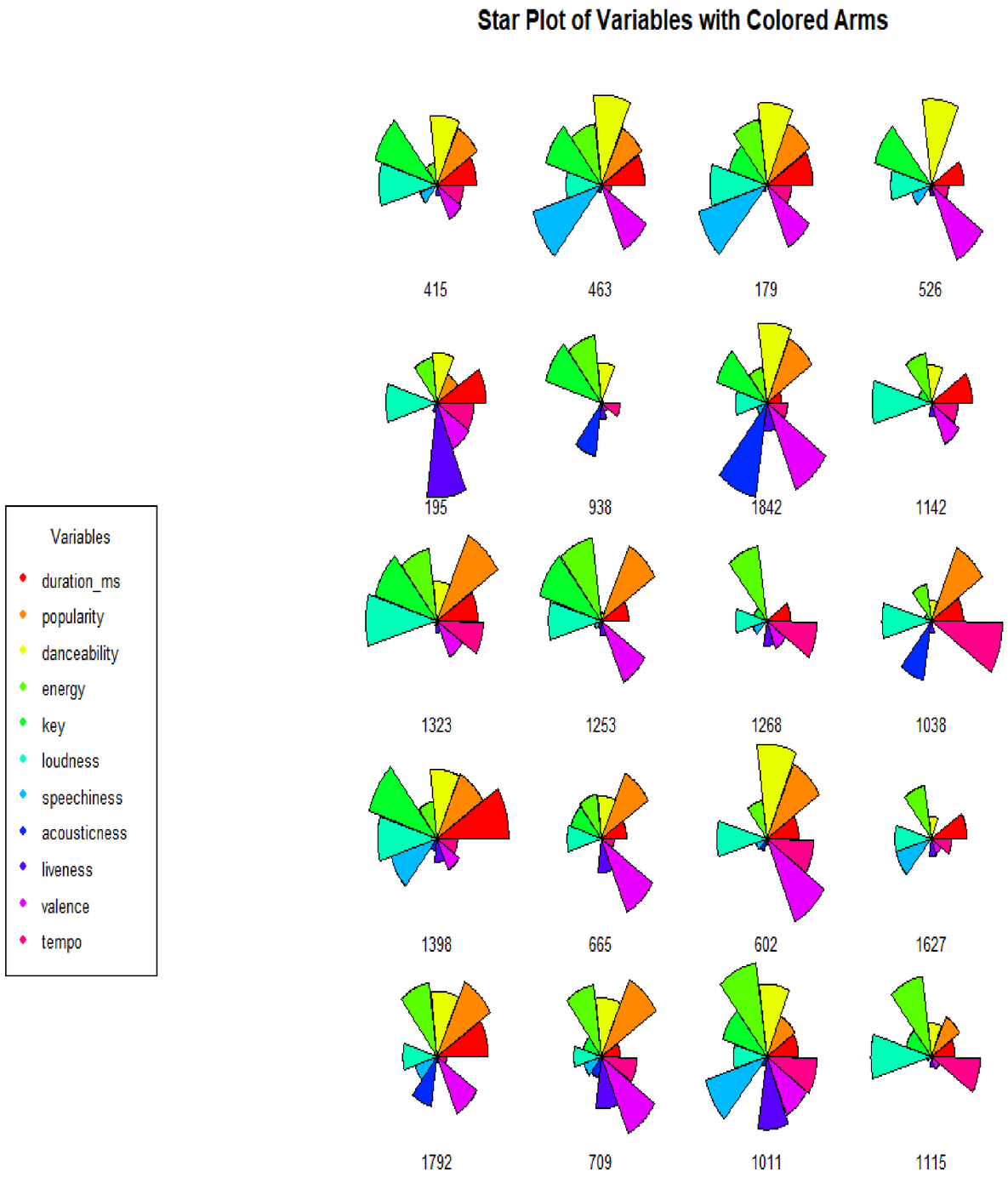


It shows scatterplots of features versus explicit content.

Explicit and non-explicit songs show no clear trend on duration of the song. Also same results can make for tempo, valence.

For loudness, explicit songs are often louder than the non-explicit songs. Same as true for energy.

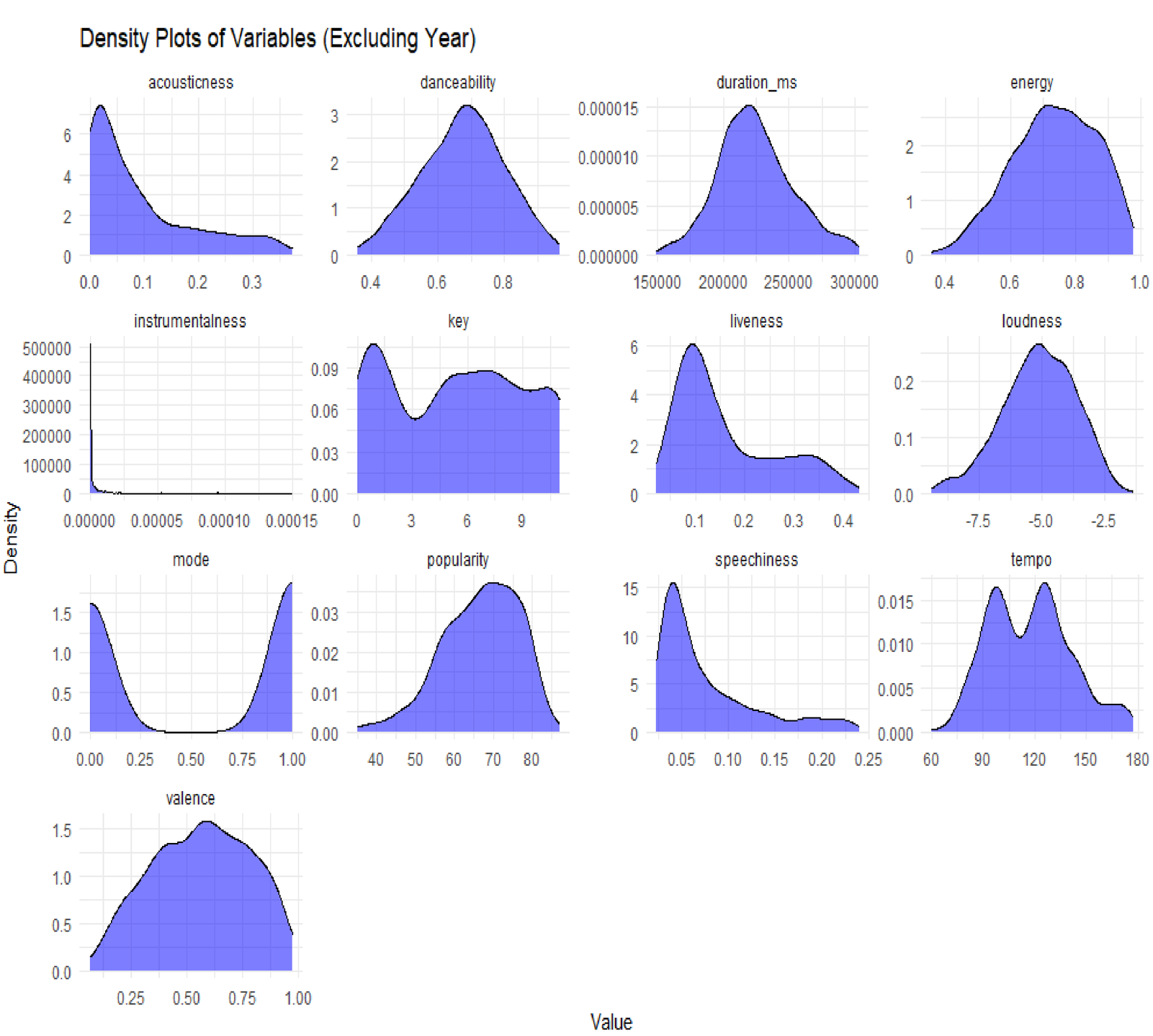
For acousticness, non-explicit songs show higher acousticness on average. While explict songs are less acoustic.



It visualizes multiple variables for individual observations in a dataset, where arm represents variables. Below the stars we can see the observation index number.

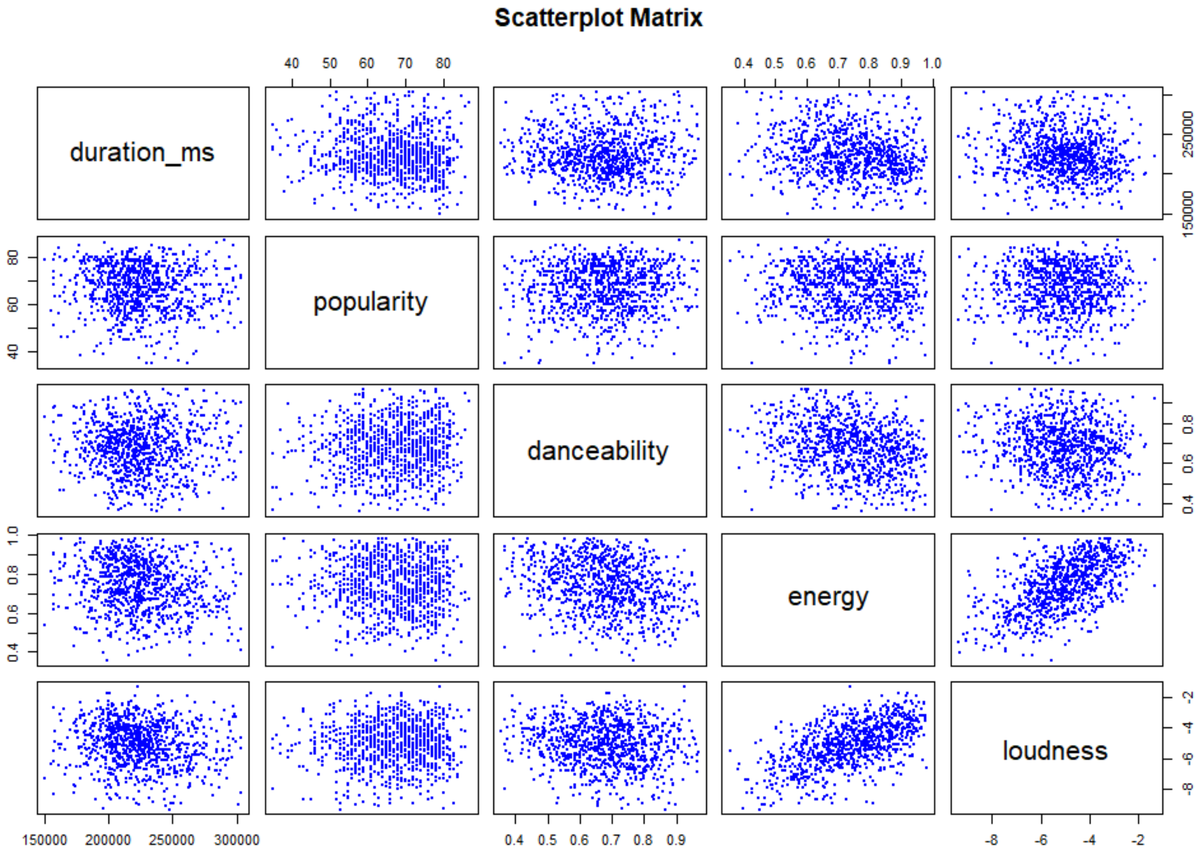
For example for observation of 415, orange are quite long, it indicates this song is really popular. Danceability is also high, it was shown as green color. Speechines is short, it was shown as blue.

For observation 1398, popularity (orange), energy (green), loudness (cyan) is high. As seen as long segment. But, tempo has short segment.



It displays density plots for various variables.

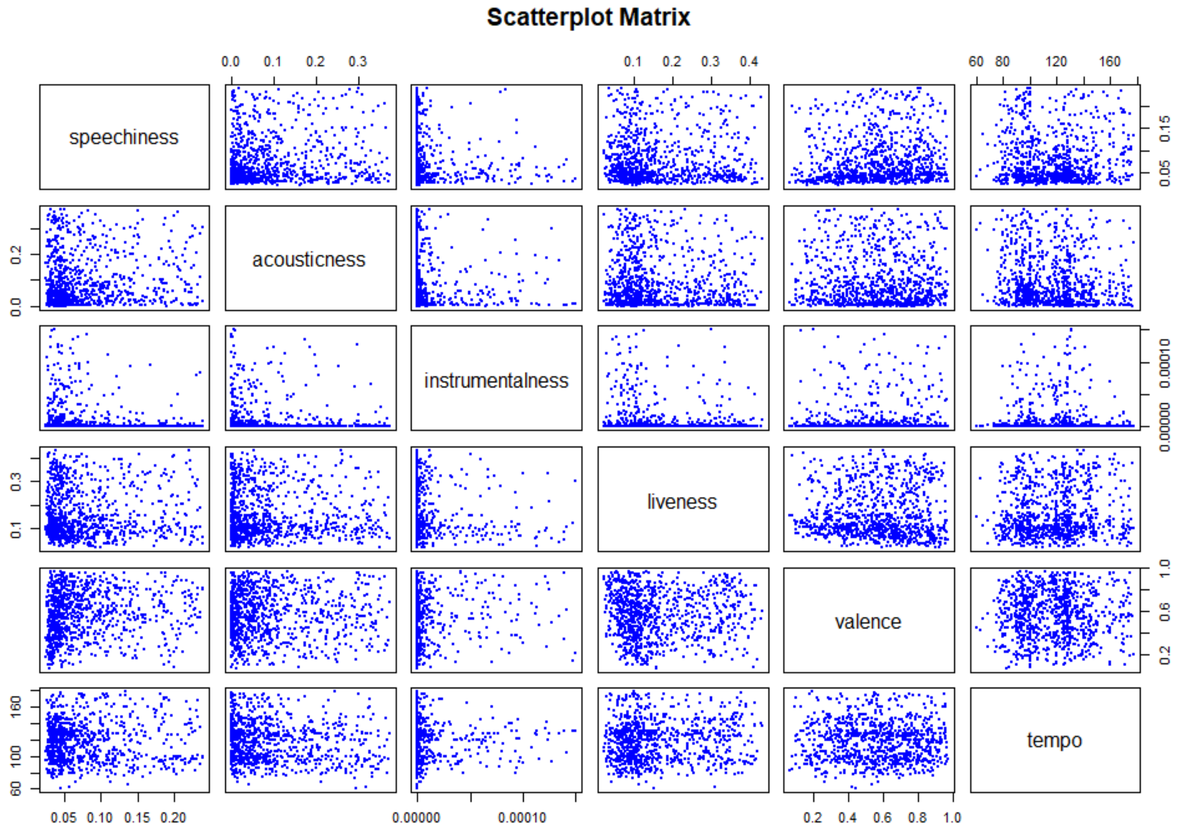
- For popularity, it's peak between range of 50 and 75, meaning most songs have moderate popularity.
- For tempo, most observations are in the peak. Same as true for loudness.
- For mode, distributed rougly around 0 and 1.
- For instrumentalness, density is highly skewed toward 0. It says no instrument.



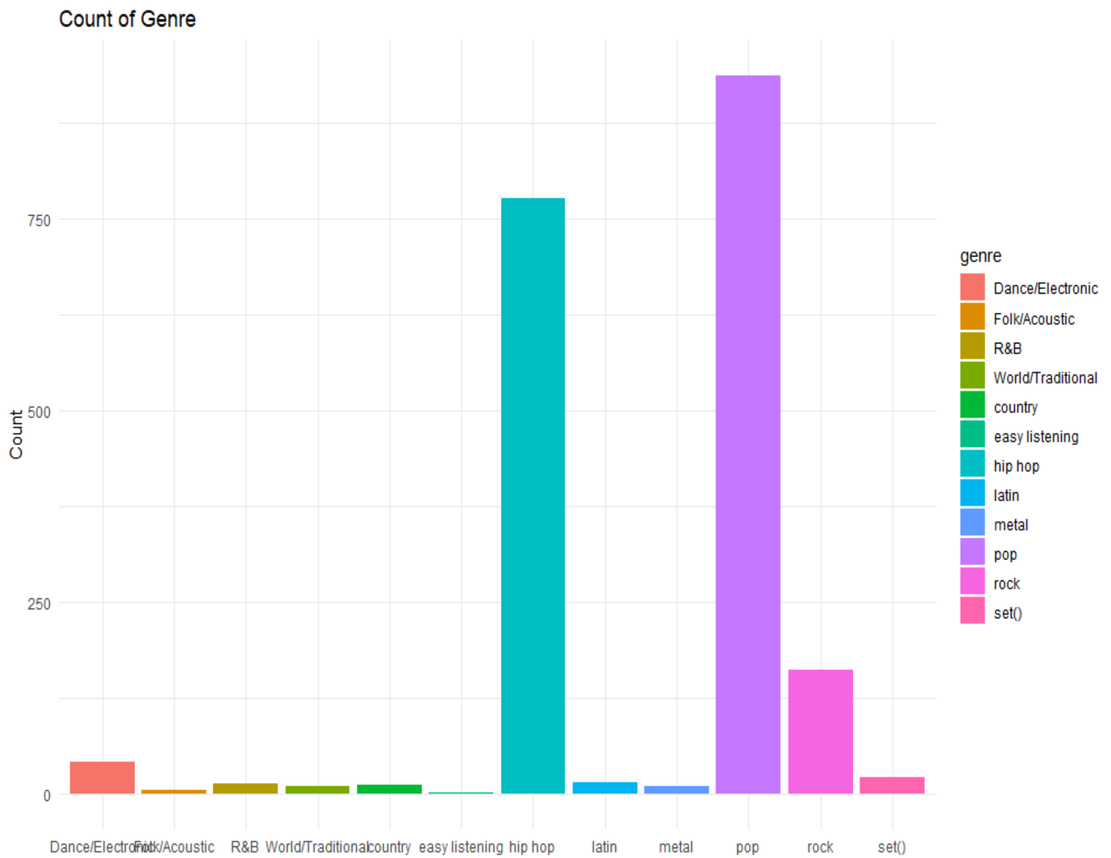
It visualizes pairwise relationships between variables.

- For duration vs other variable, no trend. Points are evenly scattered.
- For energy vs loudness gives positive relations. Louder songs tend to have higher energy levels.

- For liveness vs other graphs don't show any trend.
- For acousticness vs instrumentalness shows slightly positive trend.
- For valence vs tempo has slight cluestering in the higher tempo. It indicates more upbeat songs might have faster tempos.

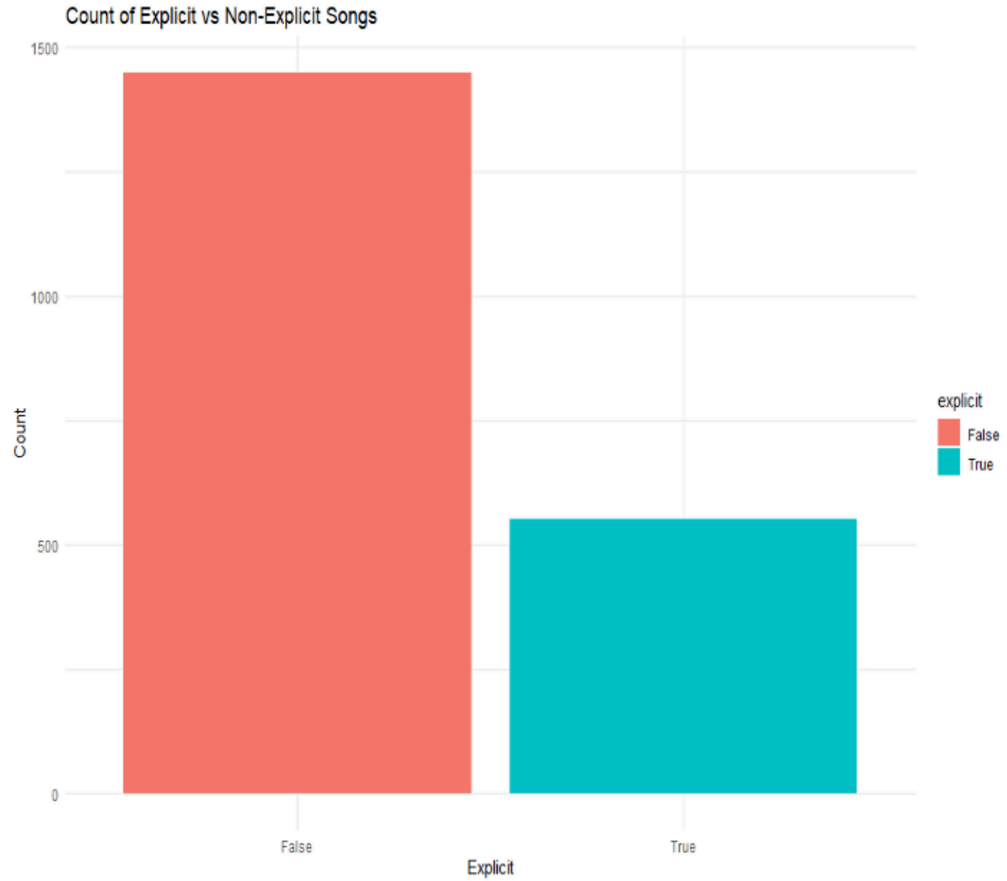


Plots for Non-numeric Variables



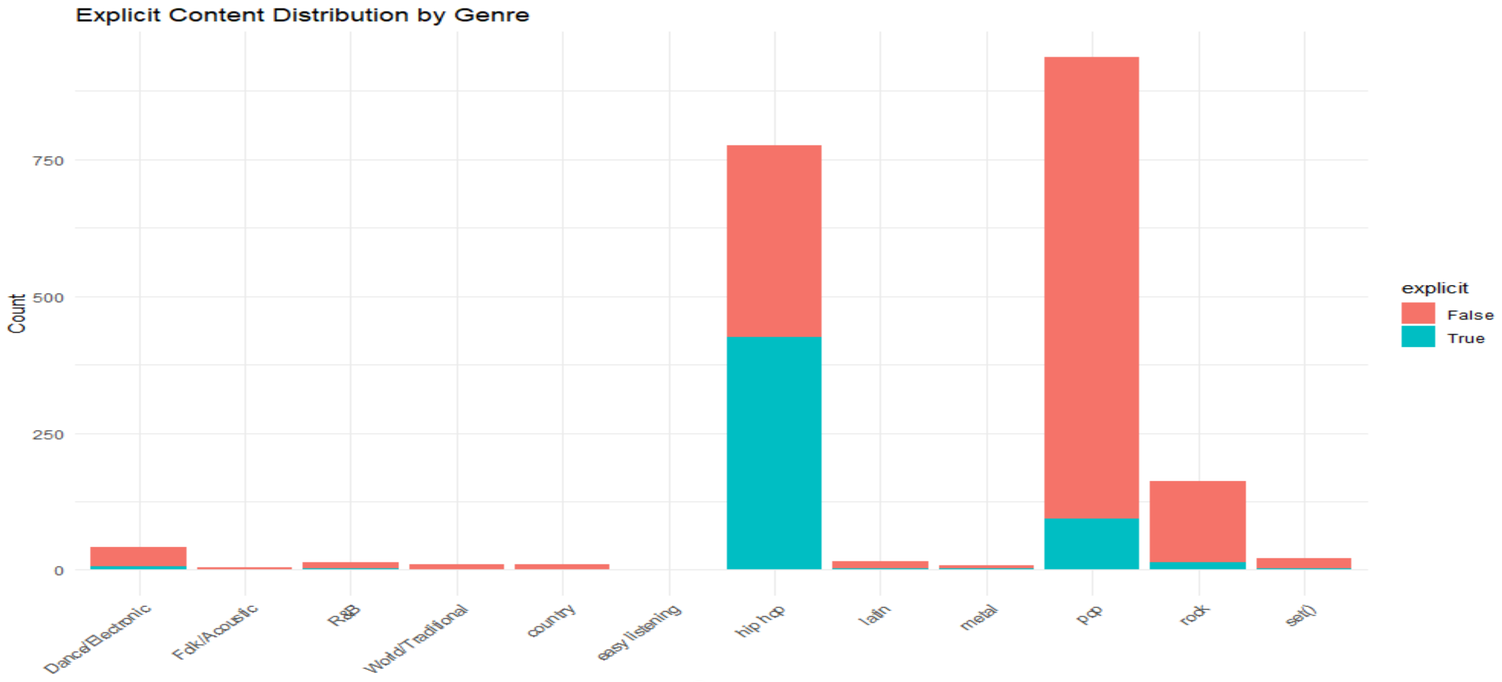
This plot shows count of songs grouped by their genre.

- Most songs are pop and hip hop type.
- Rock is the third most common genre.
- Other genres have less counts.



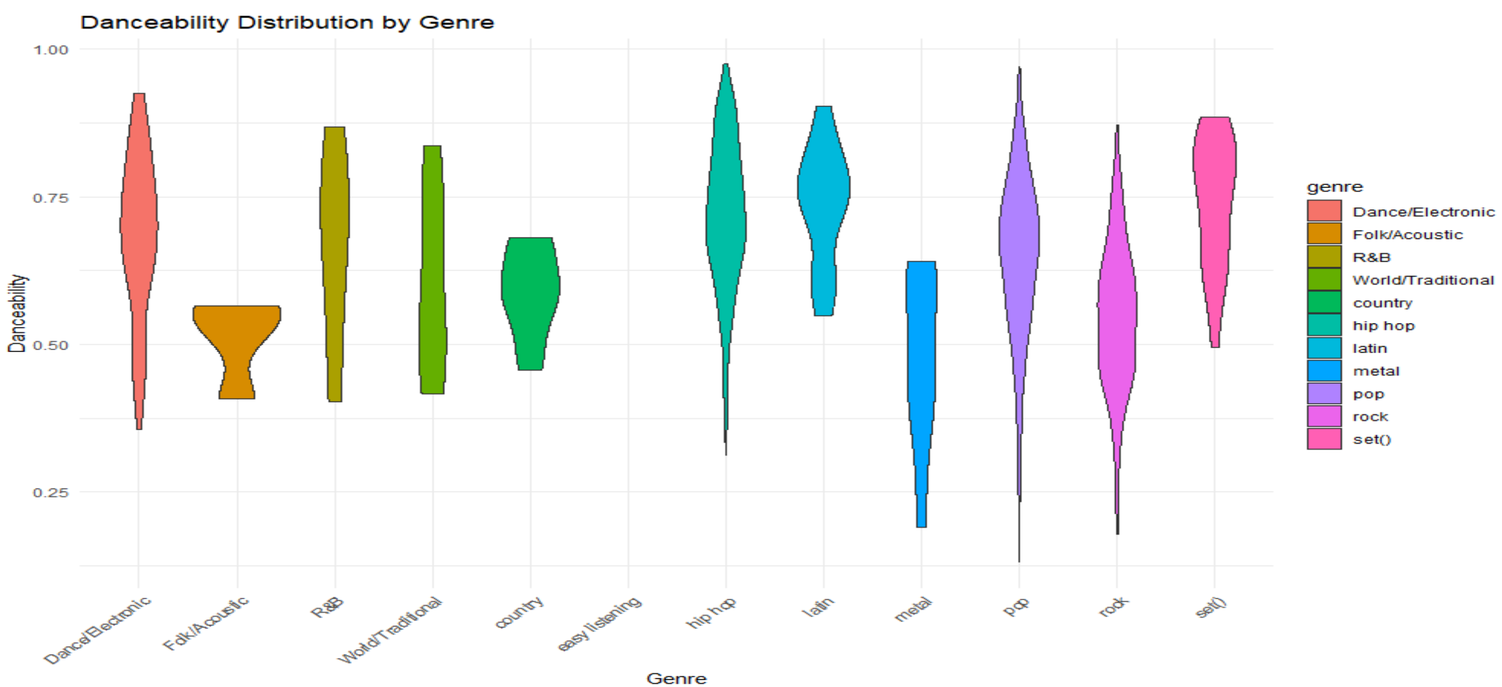
Majority of songs in the dataset are non-explicit songs.

So, many songs are suitable for children and family. These songs are labeled as family-friend songs.



It shows the distribution of explicit for each genre.

- Pop mainly has non-explicit songs.
- Hip-hop has balance between explicit and non-explicit songs.
- Rock is primarily non-explicit. This comment true for also other variables.



This plot shows realitionships between genre and danceability score.

- Electronic has a high danceability. Also this true for R&B and hip-hop.
- Rock has lower danceability.
- Metal has the lowest danceability score among variables.

Normalize Transformation for the Data

Variable	P Value	Normal or Not
acousticness	<0.05	No
danceability	<0.05	No
duration_ms	<0.05	No
energy	<0.05	No
instrumentalness	<0.05	No
key	<0.05	No
liveness	<0.05	No
loudness	<0.05	No
mode	<0.05	No
popularity	<0.05	No
speechiness	<0.05	No
tempo	<0.05	No
valence	<0.05	No

by using bestNormalize orderNorm was used

Variable	P Value	Normal Distribution
Acousticness	0.82	Yes
Danceability	<0.05	No
Duration	0.78	Yes
Energy	0.93	Yes
Instrumentalness	<0.05	No
Key	<0.05	No
Liveness	0.91	Yes
Loudness	<0.05	No
Mode	<0.05	No
Popularity	<0.05	No
Speechiness	0.76	Yes
Tempo	0.92	Yes
Valence	0.83	Yes

For future analysis especially for hypothesis testing data is separated into two parts (normal and non-normal data). Analyses were made based on normal and non-normal conditions.

Normal Variable Hotelling T^2

To check normality Royston Test is used: p-val=0.76 so normallity assumption is valid. Use Hotelling T^2.

To gain initial understanding of the normal data, Hotelling T^2 was conducted. The goal was to determine whether the column means of the normal variables are equal to hypothesized values. Based on the result below, p-value is 0.87, which is greater than 0.05. So, fail to reject H0. True mean vector is equal to the hypothesized values for each variable.

Hotelling's one sample T2-test

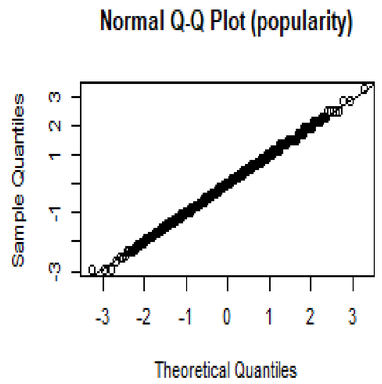
data: data_normal

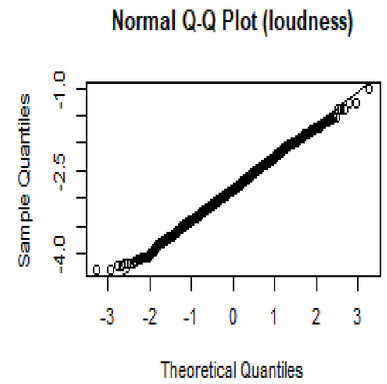
T.2 = 0, df1 = 7, df2 = 948, p-value = 0.8697531

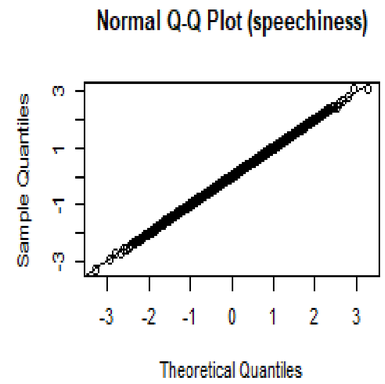
alternative hypothesis: true location is not equal to c (-0.0001884794692393 9, -2.83210081313214,

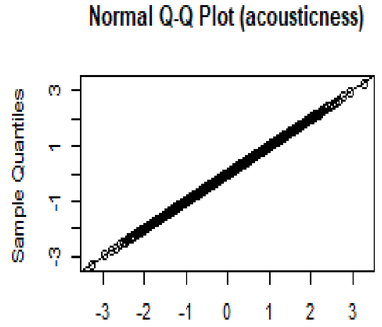
-0.0000774518747130771, -0.00000754805671991374, -0.0000052176161672183,

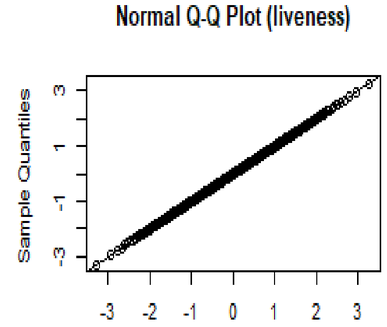
-0.0000190050401284396,0.00000624190291223819)

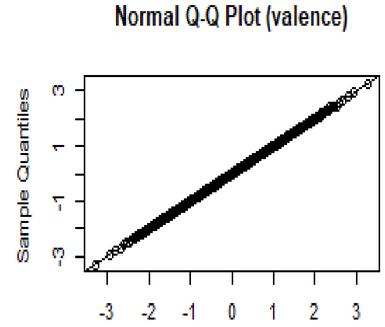


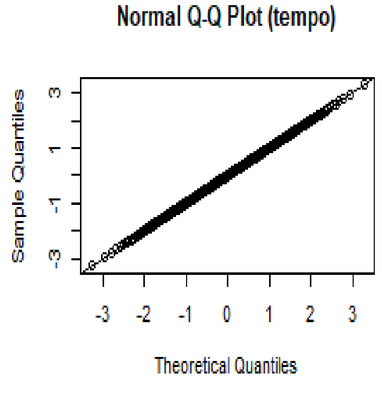








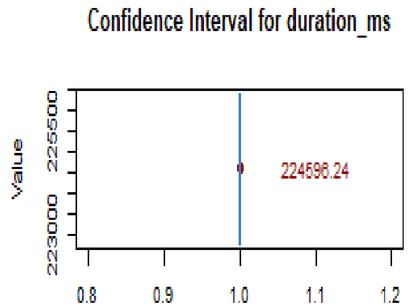


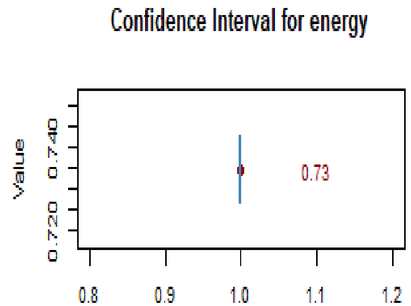


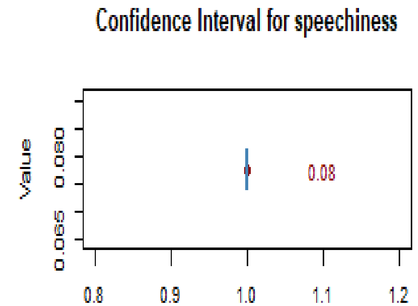
Confidence Interval For Normal Data

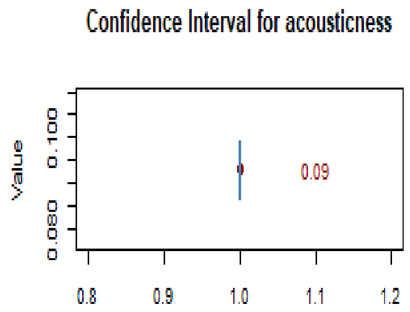
Variable	Mean	Lower CI	Upper CI	CI Range
duration	224596,244	222801,9063	226370,0159	1784,054777
energy	0,734289005	0,726622565	0,742449424	0,007913429
speechiness	0,077565759	0,074253458	0,081135471	0,003441007
acousticness	0,093143643	0,08655182	0,099325734	0,006386957
liveness	0,161187539	0,154856102	0,167635634	0,006389766
valence	0,563431728	0,548476694	0,577914188	0,014718747
tempo	117,4001613	115,9016126	118,8791083	1,488747827

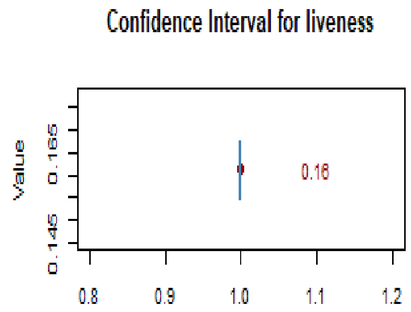
Note: These confidence interval results were obtained without scaling the variables.

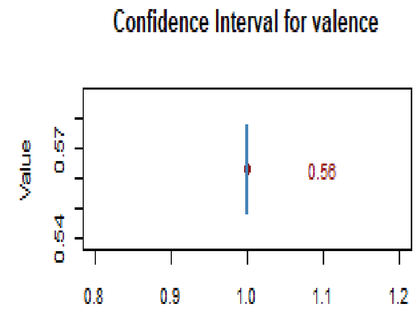


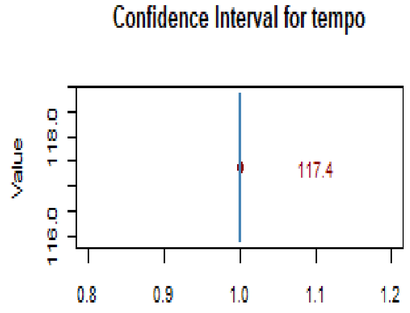












RQ : Is there a relationship between Genre and other normal data variables?

Response duration					
	df	sum sq	mean sq	F-value	p-value
Genre	11	41,45	3,7682	3,8274	<0.05
Residuals	1988	1957,24	0,9845		
Response energy					
	df	sum sq	mean sq	F-value	p-value
Genre	11	84,43	7,6757	7,9718	<0.05
Residuals	1988	1914,15	0,9629		
Response speechiness					
	df	sum sq	mean sq	F-value	p-value
Genre	11	373,45	33,95	41,53	<0.05
Residuals	1988	1625,13	0,817		
Response acousticness					
	df	sum sq	mean sq	F-value	p-value
Genre	11	106,17	9,6519	10,139	<0.05
Residuals	1988	1892,49	0,952		
Response liveness					
	df	sum sq	mean sq	F-value	p-value
Genre	11	11,23	1,0211	1,0214	0,45
Residuals	1988	1987,41	0,999		
Response valence					
	df	sum sq	mean sq	F-value	p-value
Genre	11	19,28	1,752	1,761	0,056
Residuals	1988	1978,65	0,995		
Response tempo					
	df	sum sq	mean sq	F-value	p-value
Genre	11	30,3	2,755	2,782	<0.05
Residuals	1928	1968,4	0,99		

-	df	Pillai	Approx-F	num-df	den-df	Pr
Genre	11	0,32	8,62	77	13916	<0.05
Residuals	1988					

So, there is a significant relationship between genre and other variables. To see which variables create these differences, pairwise t-test will be used as a post-hoc analysis.

Genre significantly impacts duration, energy, speechiness, acousticness, and tempo. However, it does not have a significant impact on liveness or valence based on post-hoc results.

RQ : Is there a relationship between Explicit and other normal data variables?

Response duration					
	df	sum sq	mean sq	F-value	Pr
Explicit	1	28,2	28,1994	28,593	<0,05
Residuals	1998	1970,5	0,9862		
Response energy					
	df	sum sq	mean sq	F-value	Pr
Explicit	1	55,1	55,104	56,65	<0,05
Residuals	1998	1943,5	0,973		
Response speechiness					
	df	sum sq	mean sq	F-value	Pr
Explicit	1	310,37	310,367	367,32	<0,05
Residuals	1998	1688,21	0,845		
Response acousticness					
	df	sum sq	mean sq	F-value	Pr
Explicit	1	2,59	2,5865	2,589	0,1078
Residuals	1998	1996,08	0,1		
Response liveness					
	df	sum sq	mean sq	F-value	Pr
Explicit	1	0,02	0,0175	0,0175	0,895
Residuals	1998	1998,63	1,0003		
Response valence					
	df	sum sq	mean sq	F-value	Pr
Explicit	1	4,48	4,477	4,4876	0,034
Residuals	1998	1993,45	0,998		
Response tempo					
	df	sum sq	mean sq	F-value	Pr
Explicit	1	0,01	0,0136	0,0136	0,9073
Residuals	1998	1998,66	1,0003		

-	approx-F	den-DF	Df	num-Df	Pillai	Pr(>F)
explicit	72878	1992	1	7	20388	<0.05
Residuals			1998			

So, there is a significant relationship between explicit (True and False) and other variables. To see which variables create these differences, pairwise t-test will be used as a post-hoc analysis.

Explicit significantly impacts duration, energy, valence and speechiness. However, it does not have a significant impact on accousticness, liveness, and tempo based on post-hoc results.

Multivariate Linear Regression

For further analysis, mlr was conducted. The dependent variables (y) were taken as significant for both Genre and Explicit. The independent variables (x) were taken as Genre and Explicit.

Duration = 0.12 + 0.53x(Genre Folk/Acoustic) + (-0.12)x(Genre R&B)
+ (-0.45)x(Genre World/Traditional) + 0.35x(Genre Country) + 1.59x(Genre Easy Listening)
+ 0.07x(Genre Hip Hop) + 0.24x(Genre Latin) + (-0.27)x(Genre Metal)
+ (-0.24)x(Genre Pop) + (-0.29)x(Genre Rock) + (-0.06)x(Genre Set()) + 0.18x(Explicit True)

Residual standard error = 0.99
Multiple R-squared = 0.03
Adjusted R-squared = 0.02
F-statistic = 4.33 on 12 and 1987 DF
P-value = <0.05

Energy = -0.46 + (-0.51)x(Genre Folk/Acoustic) + (-0.66)x(Genre R&B)
+ (-0.64)x(Genre World/Traditional) + (-0.10)x(Genre Country)
+ 0.11x(Genre Easy Listening) + 0.44x(Genre Hip Hop) + (-0.24)x(Genre Latin)
+ (-0.67)x(Genre Metal) + 0.09x(Genre Set()) + (-0.30)x(Genre Rock)
+ (-0.17)x(Explicit True)

Residual standard error = 0.9741
Multiple R-squared = 0.06
Adjusted R-squared = 0.051
F-statistic = 9.926 on 12 and 1987 DF
P-value = <0.05

Speechiness = -0.43 + (-0.78)x(Genre Folk/Acoustic) + (-0.41)x(Genre R&B)
+ (-0.03)x(Genre World/Traditional) + 0.02x(Genre Country) + 0.94x(Genre Easy Listening)
+ 0.76x(Genre Hip Hop) + 0.16x(Genre Latin) + (-0.11)x(Genre Metal) + 0.09x(Genre Pop)
+ 0.06x(Genre Rock) + 0.07x(Genre Set()) + 0.57x(Explicit True)

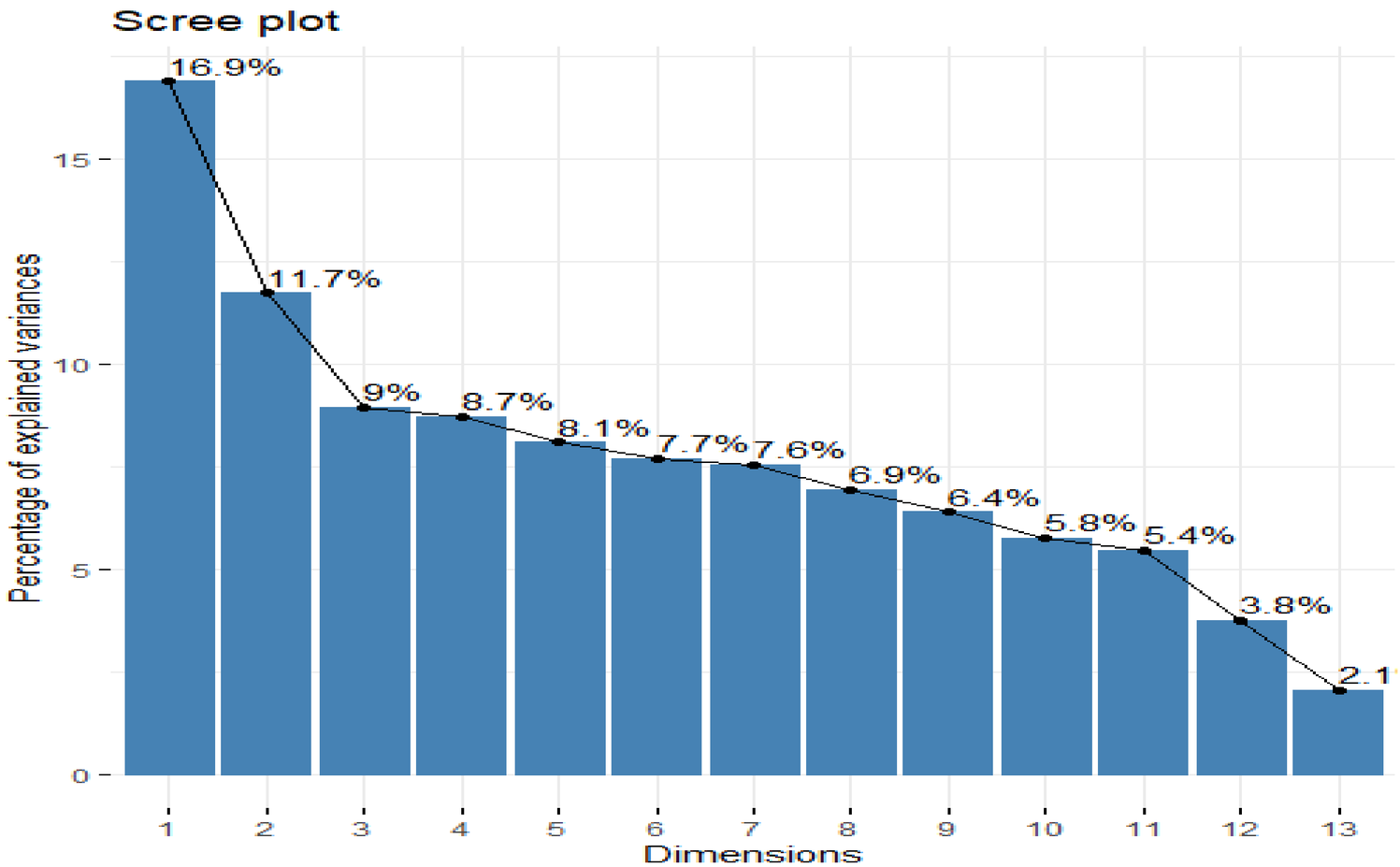
Residual standard error = 0.8778
Multiple R-squared = 0.2339
Adjusted R-squared = 0.2293
F-statistic = 50.56 on 12 and 1987 DF
P-value = <0.05

Valence = −0.077+ (−0.32)x(Genre Folk/Acoustic) + 0.45x(Genre R&B)
+ 0.56(Genre World/Traditional) + 0.26x(Genre Country)
+ (−0.55)x(Genre Easy Listening) + 0.20x(Genre Hip Hop) + 0.21x(Genre Latin)
+ (−0.49)x(Genre Metal) + 0.57x(Genre Set()) +0.08x(Genre Pop) + (−0.07)x(Genre Rock)
+(−0.17)x(Explicit True)

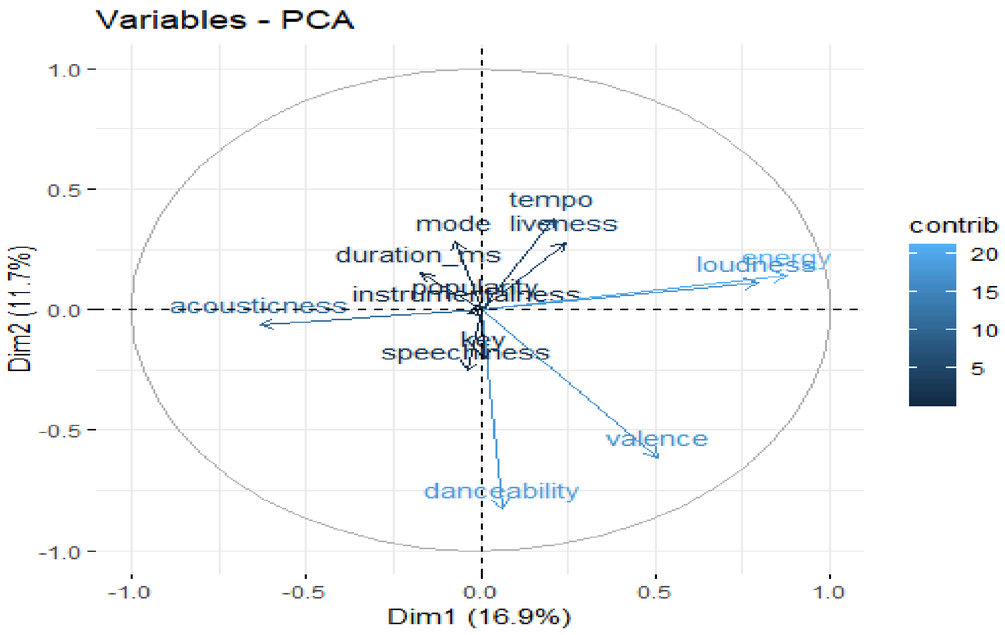
Residual standard error = 0,9956
Multiple R-squared = 0,01422
Adjusted R-squared = 0.008
F-statistic = 2,388 on 12 and 1987 DF
P-value = <0.05

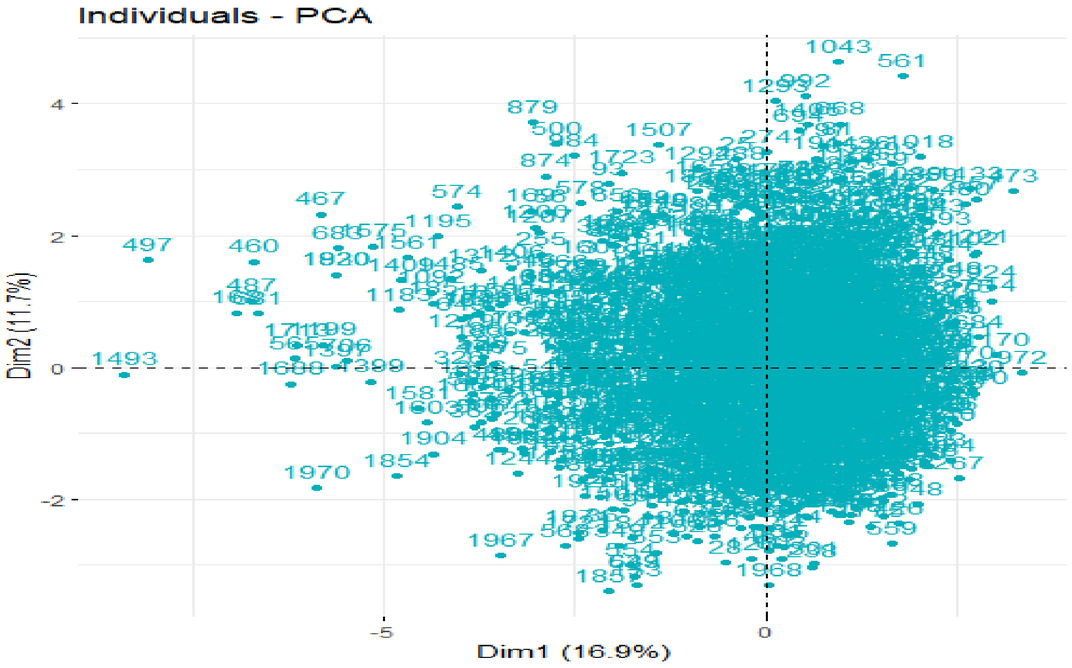
Principal Components Analysis (PCA)

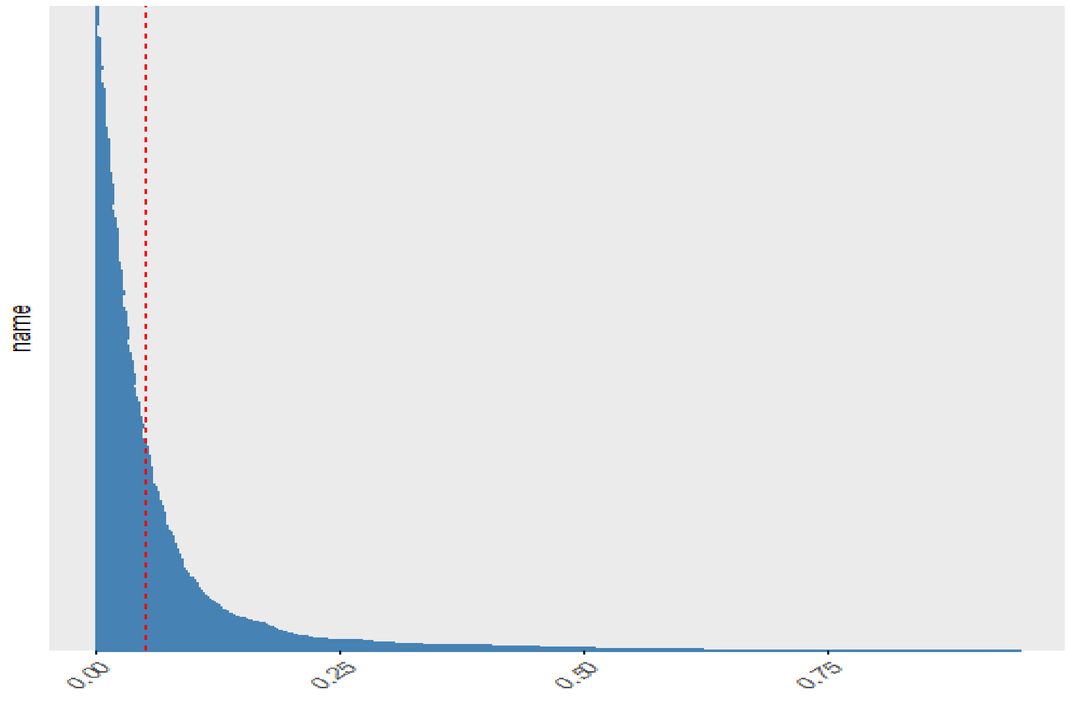
Importance of Components													
	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9	PC10	PC11	PC12	PC13
Standard deviation	1,4814	1,2349	1,0789	1,0646	1,0258	1,0001	0,9908	0,9496	0,911	0,866	0,8416	0,7002	0,5176
Proportion of Variance	0,1688	0,1173	0,0895	0,0872	0,081	0,077	0,0755	0,0755	0,0639	0,0577	0,0545	0,0377	0,0206
Cumulative Proportion	0,1688	0,2861	0,3757	0,4629	0,5438	0,6207	0,6962	0,6962	0,8295	0,8872	0,9417	0,9794	1



The PCA results gave us 13 components. However, since our data does not have multicollinearity problem, or significant outlier problem using PCA might not be meaningful. The Scree Plot does not show sharply drops off, so none of the components seem significant. We could consider first 10 components since they explain 89% of the data. But, overall PCA does not seem to add much value for our data.







PC1 explains the largest proportion of variance in the data (16.9%).
PC2 explains the second largest proportion of variance in the data (11.7%).

- Energy and loudness positively contribute to PC1.
- Acousticness negatively contribute to PC1.
- Tempo and mode are near perpendicular to other variables. So, they are independent of other variables (like energy).
- Danceability, liveness and valence have an impact on PC2.

These two graph also show us our PCA results are not revealing meaningful separation in the data. So, PCA might not be the best dimensionality reduction techniques for our data.

PCA Regression

Valence = 0.55 + (-0.02)x(PC1) + (-0.03)x(PC2) + (-0.01)x(PC3) + (-0.01)x(PC4) + 0.01x(PC5) + (-0.02)x(PC6) + (-0.03)x(PC7)

Residual standard error = 0.20

Multiple R-squared = 0.18

Adjusted R-squared = 0.18

F-statistic = 63.45 on 7 and 1992 DF

P-value = <0.05

Speechiness = 0.103 + 0.031x(PC1) + (-0.003)x(PC2) + 0.002x(PC3) + (-0.001)x(PC4) + 0.002x(PC5) + (-0.003)x(PC6) + (-0.02)x(PC7)

Residual standard error = 0.09

Multiple R-squared = 0.01

Adjusted R-squared = 0.01

F-statistic = 14.42 on 7 and 1992 DF

P-value = <0.05

Energy = 0.72 + (-0.08)x(PC1) + (-0.03)x(PC2) + (-0.02)x(PC3) + 0.01x(PC4) + 0.00x(PC5) + 0.00x(PC6) + 0.09x(PC7)

Residual standard error = 0.11

Multiple R-squared = 0.50

Adjusted R-squared = 0.50

F-statistic = 285.80 on 7 and 1992 DF

P-value = <0.05

For Valence Model:

The model explains 18% of the variance in Valence, moderate.
P-value<0.05 so significant, but Principal Components' coefficients are small (PC1-PC7), indicating weak effects.

For Speechiness Model:

The model explains 1% of the variance in Speechiness, weak.
P-value<0.05 so significant, but their explanatory power is minimal.

For Energy Model:

The model explains 50% of the variance in Energy, which is strong compared to other models.
P-value<0.05 so significant, but their explanatory power is small.

Factor Analysis and Factor Rotation

Kaiser-Meyer-Olkin factor adequacy		
Call:KMO (r=cm)		
Overall MSA = 0.56		
MSA for each item =	duration_ms	0.66
	mode	0.67
	valence	0.52
	popularity	0.47
	speechiness	0.50
	tempo	0.61
	danceability	0.41
	acousticness	0.67
	energy	0.55
	instrumentalness	0.26

chi sq
2.807.174
p value
<0.05
df
66

Factor Analysis	
	objective
factors=4	0.001
factors=5	0.002
factors=6	0.047
factors=7	0.242

First we need to look at MSA value since MSA>0.5 we can run Factor Analysis on this data. Moreover, Bartletts test must be also significant. As we can see, it is also significant. Hence, Factor Analysis is considered as an appropriate technique for further analysis of the data.

To determine the number of factors sufficient to group the variables, we used the factanal function and tested the null hypothesis that the specified number of factors is adequate. For a 7 factor solution, the p-val=0.242, which is greater than 0.05. Hence, we fail to reject H0, and conclude that 7 factor is adequate.

```
Call:
factanal(x = data_numeric, factors = 7, rotation = "varimax")

Uniquenesses:
     duration_ms       popularity     danceability           energy         loudness             mode      speechiness     acousticness instrumentalness
           0.888            0.930            0.005            0.066            0.431            0.968            0.813            0.700            0.745
        liveness          valence            tempo
           0.900            0.520            0.005

Loadings:
                 Factor1 Factor2 Factor3 Factor4 Factor5 Factor6 Factor7
duration_ms                              -0.190           0.197   0.173
popularity                                                        0.256
danceability             0.944           0.277           0.122
energy           0.898  -0.162           0.290   0.111
loudness         0.674                   0.176  -0.193  -0.175   0.119
mode                                     -0.104                  -0.107
speechiness                                              0.407
acousticness    -0.523
instrumentalness                                 0.494
liveness         0.179  -0.143                           0.187
valence          0.232   0.246           0.596
tempo            0.107           0.986

                Factor1 Factor2 Factor3 Factor4 Factor5 Factor6 Factor7
SS loadings       1.636   1.026   0.996   0.600   0.319   0.309   0.143
Proportion Var    0.136   0.086   0.083   0.050   0.027   0.026   0.012
Cumulative Var    0.136   0.222   0.305   0.355   0.381   0.407   0.419

Test of the hypothesis that 7 factors are sufficient.
The chi square statistic is 4.18 on 3 degrees of freedom.
The p-value is 0.242
```

- Danceability(0.944) and energy (0.898) load strongly onto Factor 1.
- Loudness(0.647) contributes on Factor 2.
- Tempo loads (0.986) on Factor 3.
- Valence loads (0.596) on Factor 4.
- Most variables do not have loads on Factor 5-6-7 significantly.

Factor 1 explains 13% and Factor 2 explains 8.6%.

Cumulative variance is shown as 42% which is moderate for Factor Analysis.

Factor 1 Alpha Analysis:									
	raw alpha	std.alpha	G6(smc)	average_r	S/N	ase	mean	sd	meadian_r
Reliability Analysis	0,24	0,68	0,67	0,35	2,2	0,0074	-6	0,54	0,32
Factor 2 Alpha Analysis									
	raw alpha	std.alpha	G6(smc)	average_r	S/N	ase	mean	sd	meadian_r
Reliability Analysis	0,53	0,57	0,4	0,4	1,4	0,019	0,61	0,15	0,4
Factor 3 has less than 2 items.Alpha analysis skipped									
Factor 4 Alpha Analysis									
	raw alpha	std.alpha	G6(smc)	average_r	S/N	ase	mean	sd	meadian_r
Reliability Analysis	0,48	0,45	0,46	0,21	0,8	0,018	0,65	0,12	0,33
Factor 5, 6 an 7 have less than 2 items. Alpha analysis skipped									

- Factor 1's raw alpha=0.24 and standardized alpha=0.68

These values indicate moderate reliability.

- Factor 2's raw alpha=0.53 and standardized alpha=0.57

The reliability is better than Factor 1.

- Factor 4's raw alpha=0.48 and standardized alpha=0.45

Moderate reliability.

- Factor 3,5,6 and 7 were skipped for alpha analysis because they have fewer than two items.

Fisher Discriminant Analysis

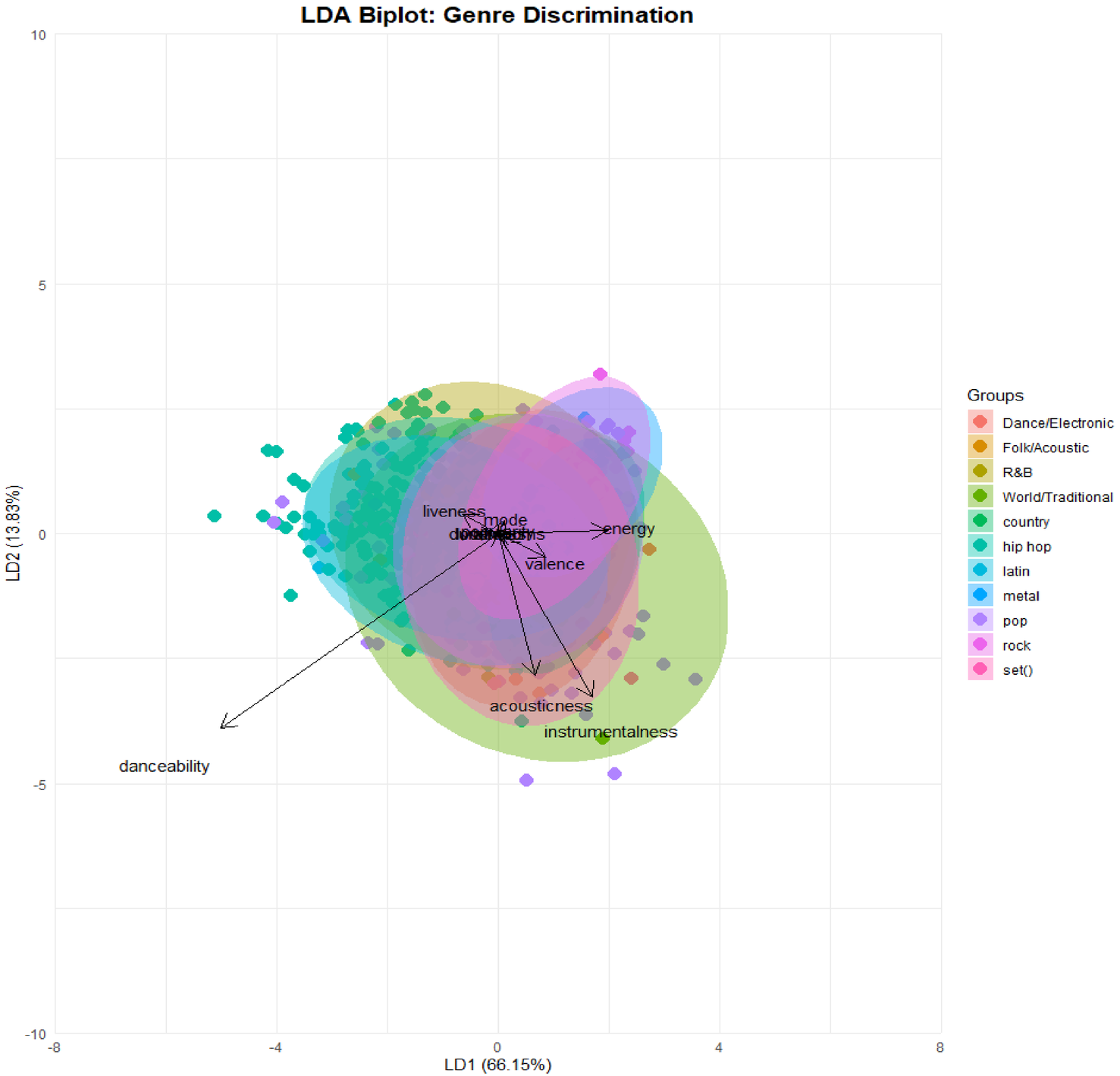
RQ : What are the most discriminative features (danceability, energy, loudness, and speechiness) for separating genres?

Coefficients of linear discriminants										
	LD1	LD2	LD3	LD4	LD5	LD6	LD7	LD8	LD9	LD10
Duration	-0,001	0,001	-0,001	0,001	-0,001	-0,001	-0,001	-0,001	-0,001	0,001
Popularity	-0,003	0,026	-0,002	0,005	0,02	-0,001	0,019	-0,005	-0,02	-0,01
Danceability	-5,017	-3,883	-1,311	-3,724	3,233	-1,574	-0,572	-0,942	-0,613	-2,332
Energy	1,974	0,062	-2,768	-3,015	6,111	0,556	-1,148	-3,487	-0,606	3,597
Loudness	-0,091	0,027	0,258	0,213	0,043	-0,187	0,001	0,45	0,28	-0,143
Mode	0,121	0,253	0,312	0,442	0,373	0,3	-0,724	-1,174	0,849	-0,896
Speechiness	-8,152	2,996	-0,414	3,316	0,949	2,518	-0,081	1,132	2,379	2,384
Acousticness	0,663	-2,846	2,74	1,736	3,823	-0,995	-1,073	-0,472	-1,879	1,191
Instrumentalness	1,705	-3,279	-7,886	7,927	1,233	-0,909	1,523	1,97	1,168	-2,912
Liveness	-0,643	0,396	-0,202	1,451	-1,05	0,522	-3,172	1,235	-3,815	-2,844
Valence	0,865	-0,482	0,68	2,185	-1,632	3,42	0,025	0,07	-0,465	1,59
Tempo	0,003	-0,003	-0,003	-0,015	0,007	0,008	-0,017	0,014	-0,007	-0,006

For LD1:
Speechiness (-8.15) strong negative impact.
Danceability (-5.01) strong negative impact.
Energy (1.97) weak positive impact.
Instrumentalness (1.7) and Acousticness (0.66) weak postive impact.

For LD2:
Speechiness (2.99) positive impact.
Acousticness (-2.85) strong negative impact.
Instrumentalness (-3.27) and Danceability (-3.88) strong negative impact.
Popularity (0.026) weak positive.

For LD3-LD10:
These axes contribute very little variance on Discriminant Analysis within specific genres.



Ellipses represent the spread of each genre in LDA. Some clusters overlap, while others are not.

- Dance/Electronic<- influenced by danceability.
- Rock<- influenced by energy and valence.
- Hip Hop<- overlaps with pop, influenced by speechiness and energy.
- Pop<- moderate contributions from most features, since it is in the center.
- World/Traditional<- influenced by liveness and acousticness
- Metal<- influenced by energy and loudness.

Dance, Folk, and Metal are more distinct. Pop, Hip Hop, R&B are overlaps.

RQ : What are the most discriminative features (danceability, energy, loudness, and speechiness) for separating Explicit (True and False)?

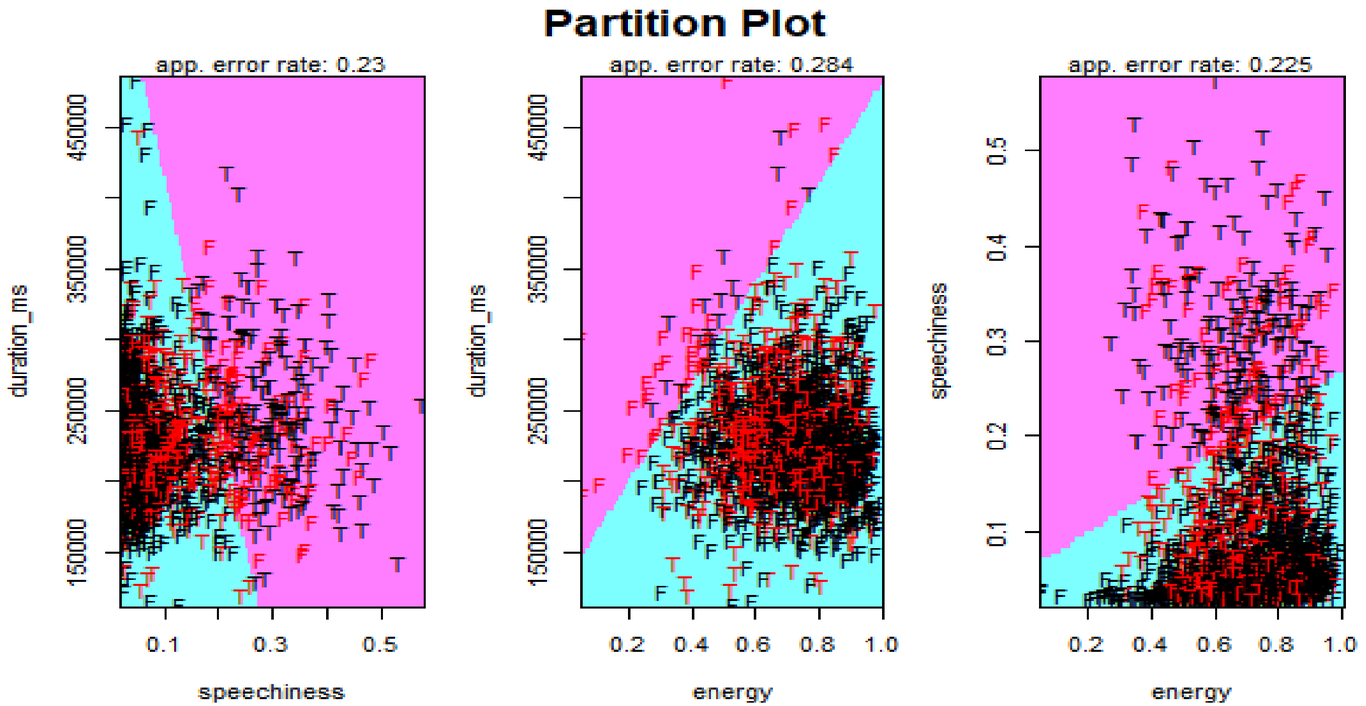
Coefficients of linear discriminants	
	LD1
Duration	0,001
Popularity	0,006
Danceability	4,357
Energy	-1,815
Loudness	0,043
Mode	0,217
Speechiness	8,8
Acousticness	-0,948
İnstrumentalness	-1,323
Liveness	0,305
Valence	-1,449
Tempo	0,004

- The most signnificant feature is Speechiness (8,8). Songs with low speechiness are more likely to be non-explicit.
- Danceability (4.36) is moderate positive. Explicit songs are more likely to have more danceability scores.
- Energy (-1.82) is weak negaitve. Explicit songs have lower energy levels.
- Valence (-1.45) and Instrumentalness (-1.32) are weak negaitve.

Strong Predictors<- Speechiness, danceability.

Negative Predictors<- Energy, valence, instrumentalness, acousticness.

Weak Predictors<- Duration, popularity, tempo.



Each panel represents combination of two features based on explicit true and false.

T= Explicit (true) (pink region)
F= Explicit (false) (blue region)
Misclassified points are shown as red point.

- Speechiness vs Duration <- High speechiness (right) points are correctly classified as explicit. Low speechiness (left) are correctly classified as non-explicit (blue region). Duration plays a minor role.
- Energy vs Liveness <- Energy has moderate power. High energy songs are mostly non-explicit. Low energy songs are mostly explicit. Liveness has a weaker influence.

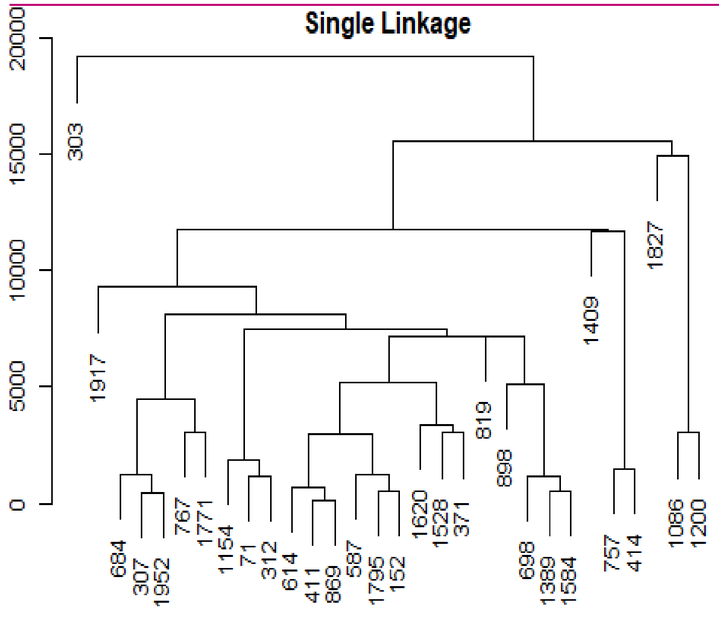
Most Important Feature<- Speechiness
Moderate Features<- Energy, duration
Weak Features<- Liveness

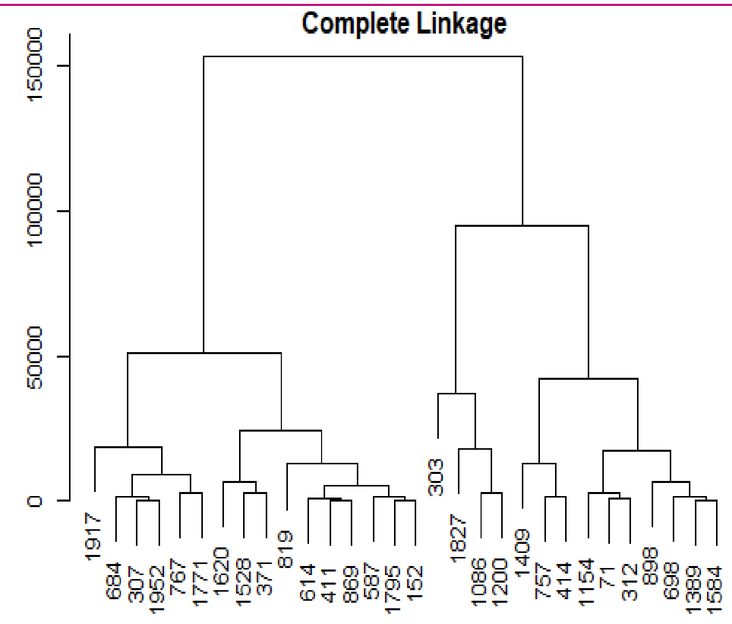
Speechiness and Energy achieves the best combinations. (lowest error rate).

CLUSTERING

Dendrograms represent hierarchical performed with different methods.

Single Linkage<- by linking the nearest point.
Complete Linkage<- by linking the farthest point.
Average Linkage<- by linking the average distance.
Ward's Method<- by minimizing the increase in total within-cluster variance.



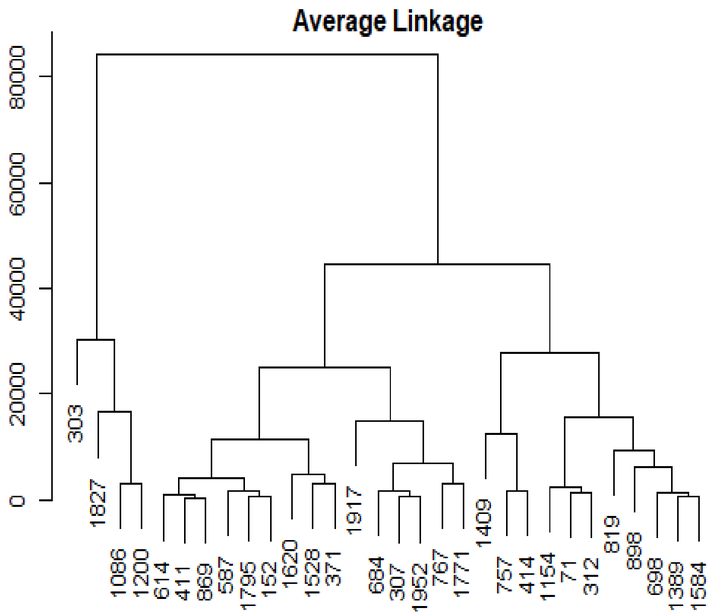


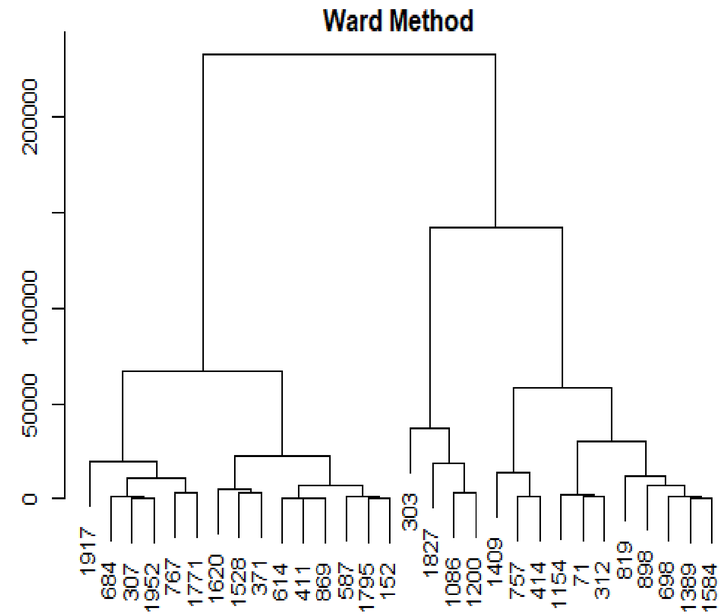
• Single Linkage
It is not ideal for our data since it fails to create compact, tidy clusters.

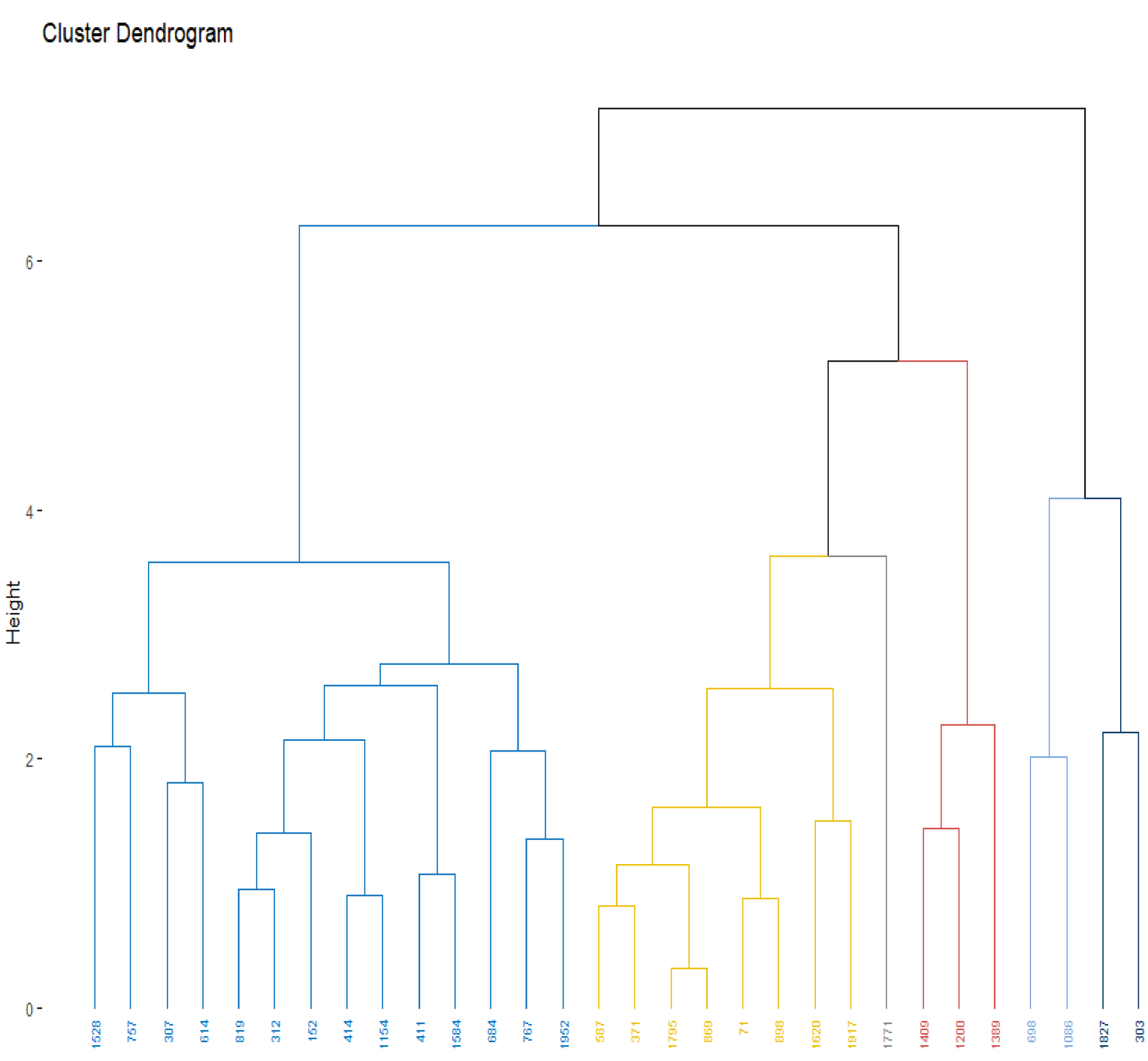
• Complete Linkage
303 and 1917 are outlier since they remain separate from other observations.
684, 307 and 1952 merge together.

• Average Linkage
684, 307, 1952 are merge together at lower. Later, 767 and 1771 are merge these observations later.

• Ward's Method
303, 1827, 1086 and 1200 are cluster together at higher distance.







Cluster 1 (Blue): 1528, 757, 307, 614, 819, 312, 152, 414, 1154, 411, 1584, 684, 767 and 1952 are in the same cluster. These observations cluster at low heights, indicates high similarity among them.

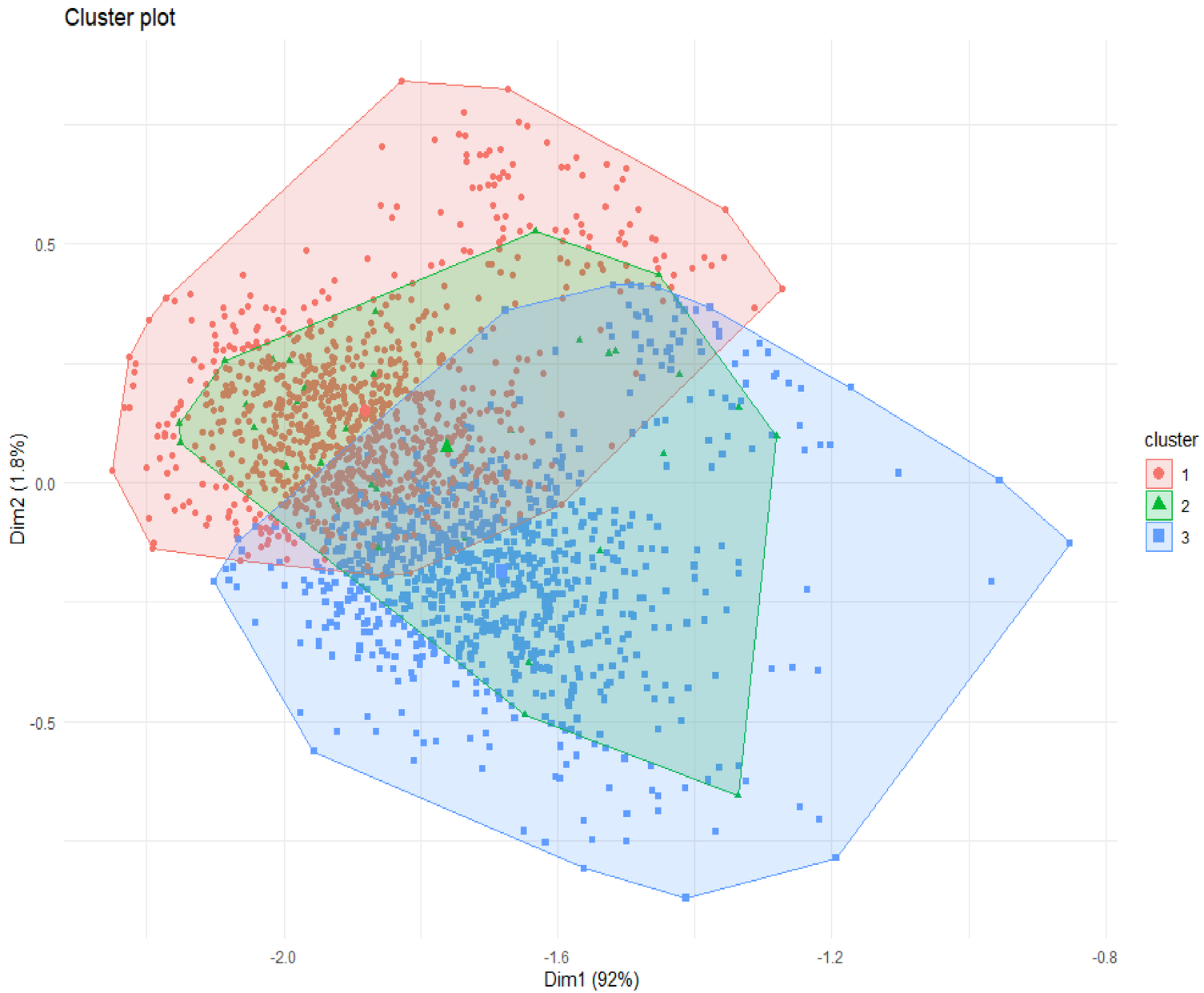
Cluster 2 (Yellow): 687, 371, 1795, 869, 71, 898, 1620 and 1917 are in the same cluster.

Cluster 3 (Red): 1409, 1200 and 1389 are in the same cluster.

Cluster 4 (Black): 1827 and 303 are in the same cluster.
Least compact since they merged at high heights.

K-Means Cluestering

Based on elbow plot we decided number of cluster as 3. So, further analysis K-means Cluestering plot was obtained for k=3 case.



Dim 2 explains 92% of the variance.
Dim 1 explains 8% of the variance.

Cluster 1 (Red): It is seperated from other significantly so this cluster has distinct characteristics.

Cluster 2 (Green): It is in the center. It overlaps with other clusters. This cluster has some level of similarity with cluster 1 and 3.

Cluster 3 (Blue): It is in the bottom-right. It has some level of significantly different characteristics with other cluster.

Canonical Correlation Analysis

Based on our research questions two groups were created.

Group 1: Significant variables for both genre and explicit.
So, duration, speechiness, energy and valence.

Group 2: Not significant for both genre and explicit.
So, danceability, tempo, popularity and acousticness.

GROUP 1

	[,1]	[,2]	[,3]	[,4]
Duration	-0,001	-0,001	0,001	-0,001
Speechiness	-0,202	-0,047	0,0165	0,577
Energy	0,182	-0,181	0,001	0,061
Valence	-0,117	-0,080	-0,007	-0,071

GROUP 2

	[,1]	[,2]	[,3]	[,4]
Danceability	-0,241	-0,095	-0,045	-0,004
Tempo	0,001	-0,001	-0,001	0,001
Popularity	0,001	0,001	-0,002	-0,001
Acousticness	-0,081	0,245	0,046	0,206

For Group 1:

Canonical Variable 1 <- Speechiness (-0.202) and energy (0.182) contribute the most. Speechiness contributes negatively. Energy contributes positively, indicates positive relationship with the CV1. Duration contributes small (no influence).

Canonical Variable 2 <- Energy (-0.181) is the most variable, contribute negatively. Speechiness (-0.047) is weak negative effect on CV2.

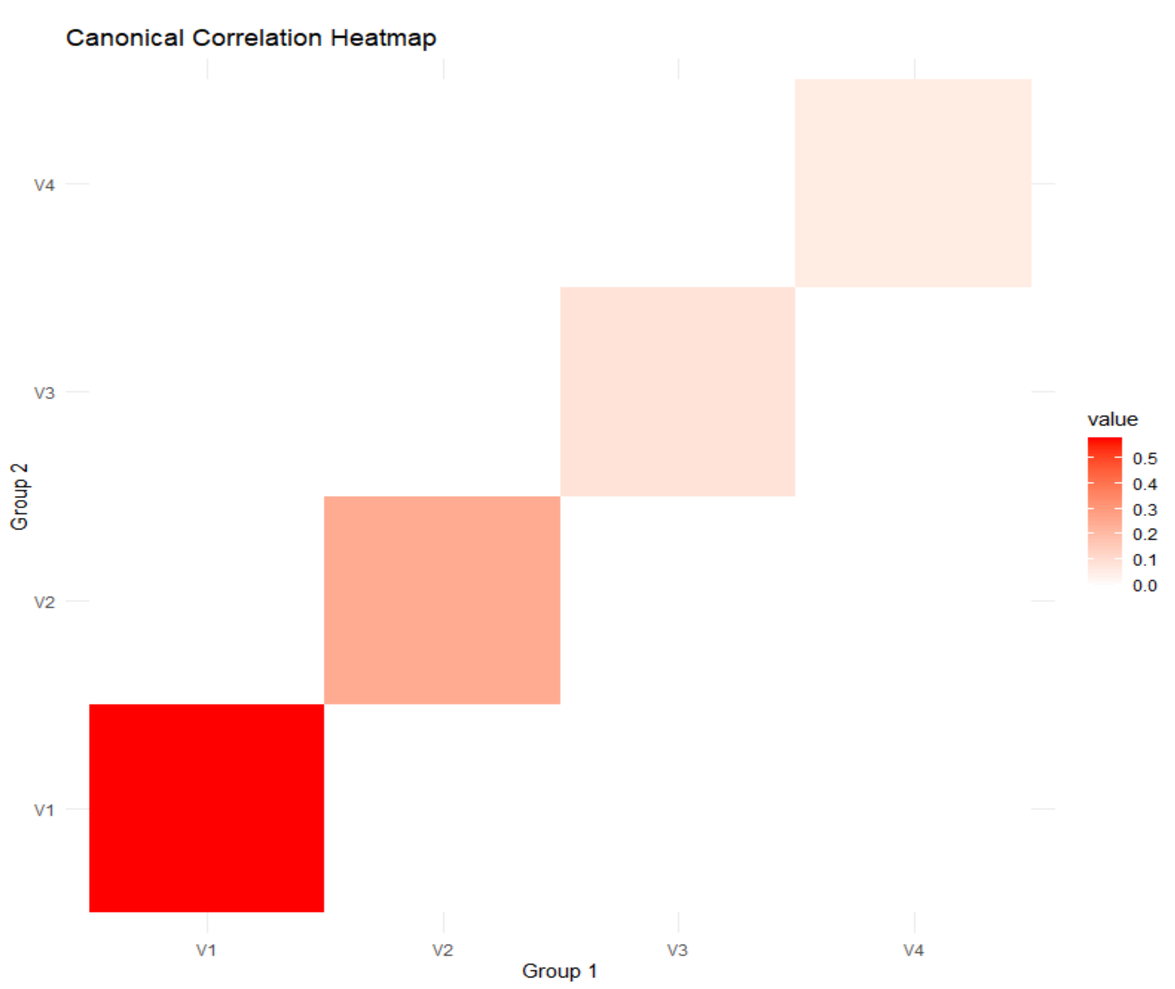
Canonical Variable 3 and 4 <- Interpretations can make similarly.

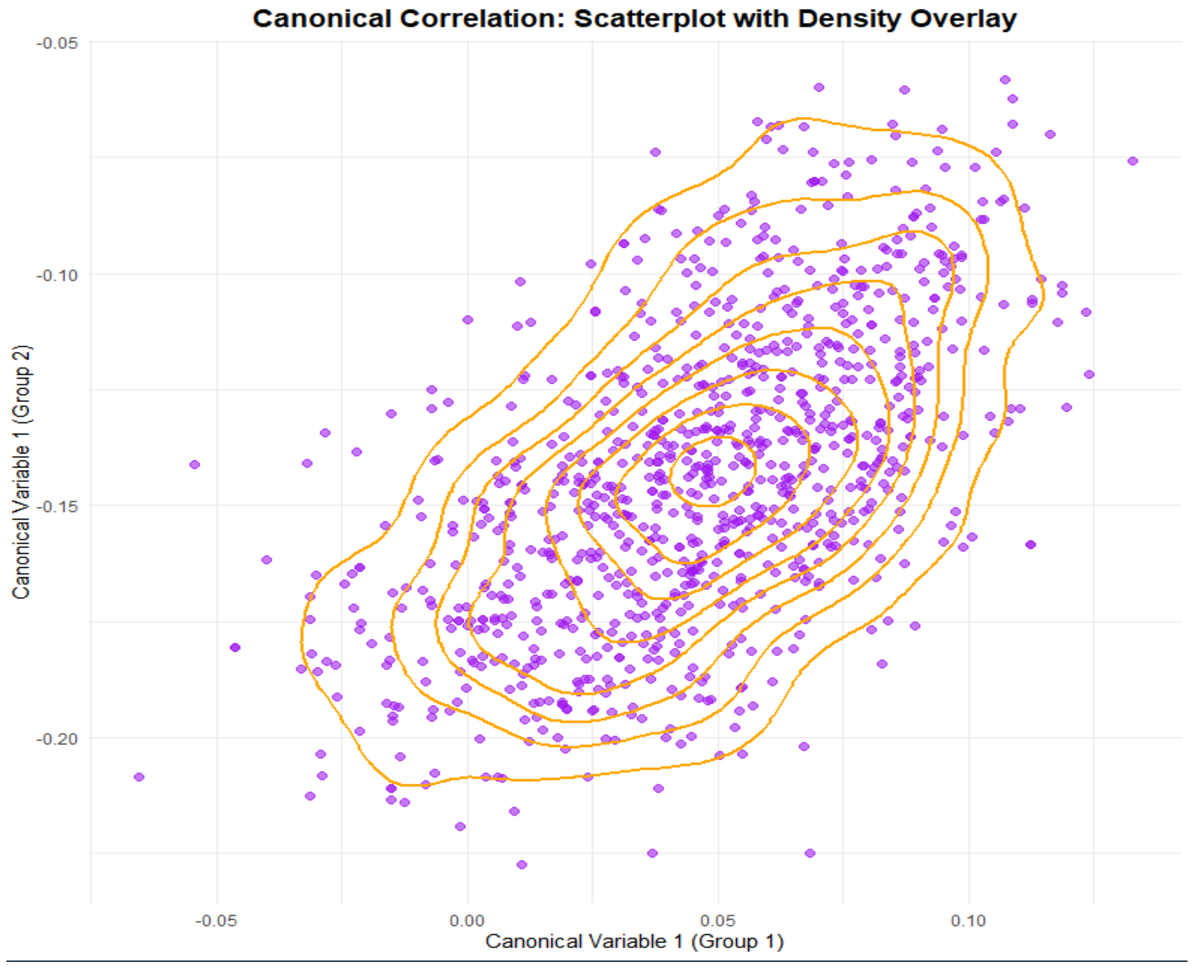
For Group 2:

Canonical Variable 1<- Danceability (-0.241) contributes most, negative impact on CV1.
Acousticness (-0.081) negative weak impacts on CV1. Tempo and popularity has weak impact.

Canonical Variable 2 <- Acousticness (0.245) contributes most, positively. Danceability has weak negative impact. Tempo and popularity have minor impact.

Canonical Variable 3 and 4 <- Interpretations can make similarly.





Each block in the heatmap represents the correlation between the respective canonical variates.
Each canonical variate from group 1 is paired with other group 2.

The most significant one is group 1 V1 and group 2 V1 since it was shown as bright red color (Strongest relationship). When we move to the up, the intensity of the red color is decreases.
For no-off diagonal correlations' blocks are white. This one is expected because correlations can occur in the same rank in CCA.

It confirms a strong and consistent relationship between the first canonical variates of group 1 and group 2 as shown as linear trend and dense clustering of points. It exhibits a clear positive linear trend, which indicates a strong relationship between group 1 and 2. As group 1 increases, so group 2 also increases. The outer contours represent observations that deviate more from the central trend. There are small number of outliers.