# Spotify Youtube

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

BY

Zarrin Yusibova 2490043
Eren Duralı 2361244

January 2024

**ABSTRACT**

This project uses multivariate analysis to explore patterns in the "Spotify and Youtube" dataset. The report provides insights into the distribution of main attributes, relationships between variables, and the impact of album types on music characteristics. The analysis includes normality checks, Hotelling's T-squared test, MANOVA, a Principal Component Regression (PCR) model, factor analysis and factor rotation, discrimination and classification, clustering, and canonical correlation analysis.

## 1. Introduction

The "Spotify and YouTube" dataset, with 26 variables per song, has insights into user interactions in the digital music platform. Having variables like danceability, energy, views, and likes, this study applies advanced statistical methods to demonstrate patterns and relationships. Exploring factors like album types, the research contributes to understanding user behavior in the platforms of Spotify and YouTube.

### 1.1. Data description

There are 26 variables associated with each song, including aspects such as track details, artist information, and URLs for both Spotify and YouTube.

### 1.2. Research questions

1) Are the means of danceability and valence significantly different from the hypothesized values of 0.7?
2) Does the type of album significantly impact the danceability, energy, liveness, and valence of the songs?
3) How effective is principal component regression in predicting the 'base_total' variable using the principal components derived from the danceability, valence, energy, and liveness variables?
4) What are the connections between different variables in the dataset? Can factor analysis provide information about the variables' relationship?
5) Are there any specific variables that have strong correlations with identified factors?

6) Is the Fisher Discriminant Analysis method effective in classifying album types? Are there any important variables or combinations for album type classification?
7) Is the Canonical correlation analysis helpful for explaining the relationship between popularity variables (Views and Likes) and song variables (danceability, energy, etc.)? How many dimensions are required to explain the correlations between these two groups?

### 1.3. Aim of the study

This study aims to conduct a multivariate analysis of the "Spotify and YouTube" dataset to show patterns and insights about user preferences in the digital music platform. Specifically, variables like danceability, energy, and album types are chosen to analyze the data deeply on user behavior and trends.

## 2. Methodology/Analysis

The methodology employs visualizations like histograms and scatter plots for data exploration in the "Spotify and YouTube" dataset. Normality checks, Hotelling's T-squared test, and MANOVA are used to examine distributional characteristics, assess means, and explore the impact of album types. Principal Component Analysis (PCA) and Principal Component Regression (PCR) are integrated for a comprehensive understanding of variable relationships and predictive capabilities within the dataset.

## 3. Results and Findings

Results include visualizations of the distribution of danceability, scatter plots of energy vs. valence, box plots of danceability by album type, and distribution of album types. It also includes normality checks and the interpretation of Hotelling's T-squared test. Additionally, the report explores the impact of album types on danceability and energy based on MANOVA results. Factor analysis explains the relations between variables via factors. Fisher discriminant analysis for album type classification. K-means clustering method for natural grouping of the variables. Multidimensional scaling for finding required dimensions. Lastly, canonical correlation analysis explains the variability of views and likes.

### 3.1 Exploratory Data Analysis

In the first output, we apply the Royston test to check multivariate normality

H0: The data follows the normal distribution.

H1: The data does not follow the normal distribution.

```
##      Test        H p value MVN
## 1 Royston 2541.242       0  NO
```

Figure 1: Royston test for multivariate normality

Since the p-value is less than alpha=0.05, we reject H0, and we can say that we do not have enough evidence to prove that the data follow the normal distribution.

### 3.2 Inferences About a Mean Vector

The second output is about the inferences about a mean vector. We choose danceability and valence variables for this one and use Hotelling's T-square test.

**H0: μ=μ0**

**H1=μ≠μ0**

```
        Hotelling's one sample T2-test

data:  y
T.2 = 514.44, df1 = 2, df2 = 1888, p-value < 2.2e-16
alternative hypothesis: true location is not equal to c(0.7,0.7)
```

Figure 2: Hotelling's T-square test for danceability and valence

Since p<alpha=0.05, we reject H0. Therefore, we do not have enough evidence to conclude that the mean vector equals (0.7,0.7).

### 3.3 Comparisons of Several Multivariate Means

```
 Response Danceability :
              Df Sum Sq Mean Sq F value    Pr(>F)
Album_type     2  1.587 0.79340   29.24 3.12e-13 ***
Residuals   1887 51.202 0.02713
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

 Response Energy :
              Df Sum Sq Mean Sq F value     Pr(>F)
Album_type     2  1.329 0.66467   14.69 4.669e-07 ***
Residuals   1887 85.378 0.04525
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

 Response Liveness :
              Df Sum Sq  Mean Sq F value Pr(>F)
Album_type     2  0.070 0.034873   1.364 0.2559
Residuals   1887 48.243 0.025566

 Response Valence :
              Df  Sum Sq  Mean Sq F value Pr(>F)
Album_type     2   0.107 0.053283  0.8888 0.4113
Residuals   1887 113.119 0.059946
```

Figure 3: MANOVA Results

Our third output contains comparisons of several multivariate means. We choose that album type as our grouping variable with three (3) categories such as “album”, “single” and “compilation”. Our independent variables were danceability, energy, liveness, and valence. We use MANOVA for this output.

Danceability and energy have significant p values. Liveness and valence p-values are not significant since alpha = 0.05.

### 3.4 Principal Component Analysis and Principal Components Regression

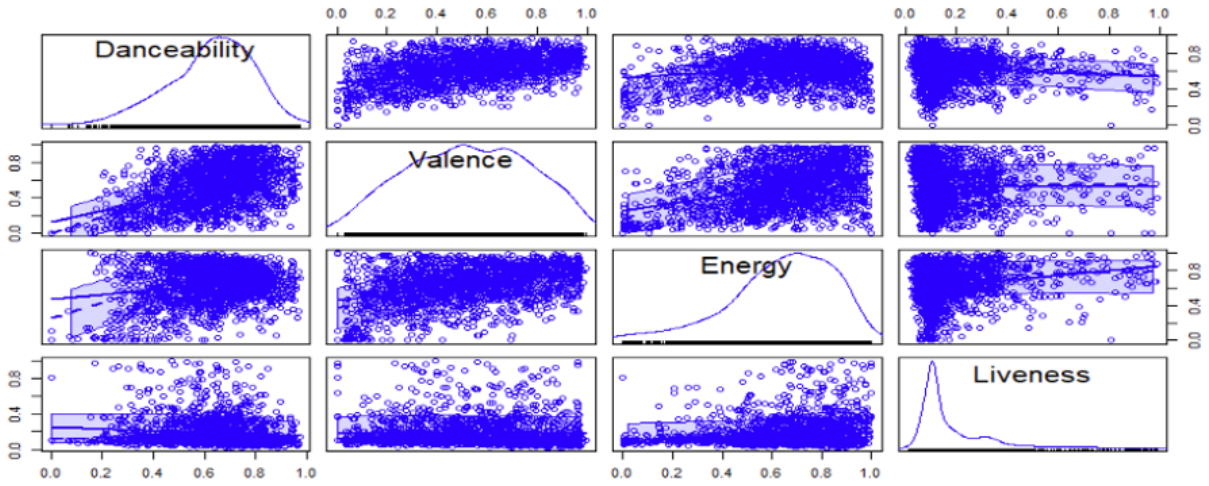



Figure 4: Scatterplot of danceability, valence, energy, and liveness

We scaled the dataset and applied the principal component analysis.

```
Importance of components:
                          PC1    PC2    PC3
Standard deviation     1.1116 1.0500 0.8136
Proportion of Variance 0.4119 0.3675 0.2206
Cumulative Proportion  0.4119 0.7794 1.0000
```

Figure 5: Importance of components

```
                    PC1         PC2         PC3
Danceability 0.4735516 -0.68949253  0.5480410
Energy       0.7674845  0.01776261 -0.6408214
Liveness     0.4321069  0.72407495  0.5375864
```

Figure 6: Rotation of PCA

We extract the first two components and continue our analysis with them.

```
Residuals:
     Min       1Q   Median       3Q      Max
-2.62190 -0.59256 -0.00188  0.61268  2.37122

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.148e-16  1.944e-02    0.00        1
PC1          4.099e-01  1.749e-02   23.43   <2e-16 ***
PC2         -2.678e-01  1.852e-02  -14.46   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.845 on 1887 degrees of freedom
Multiple R-squared:  0.2867,    Adjusted R-squared:  0.2859
F-statistic: 379.1 on 2 and 1887 DF,  p-value: < 2.2e-16
```

Figure 7: Principal Component Analysis

PC1 and PC2 are significant.

### 3.5 Factor Analysis and Factor Rotation

In the Factor analysis part, we use scaled data because factor analysis is sensitive for different scales. We did Kaiser-Meyer-Olkin factor adequacy for the test so is it reasonable to do factor analysis or not?

```
Kaiser-Meyer-Olkin factor adequacy
Call: KMO(r = cm)
Overall MSA =  0.56
MSA for each item =
      Views       Likes Danceability      Energy    Loudness Speechiness    Liveness     Valence       Tempo
Duration_ms
       0.51        0.51        0.57        0.57        0.59        0.60        0.54        0.64        0.56
0.58
```

Figure 8: Kaiser-Meyer-Olkin factor adequacy

Since the overall MSA is 0.56 > 0.5, it is reasonable to apply factor analysis.

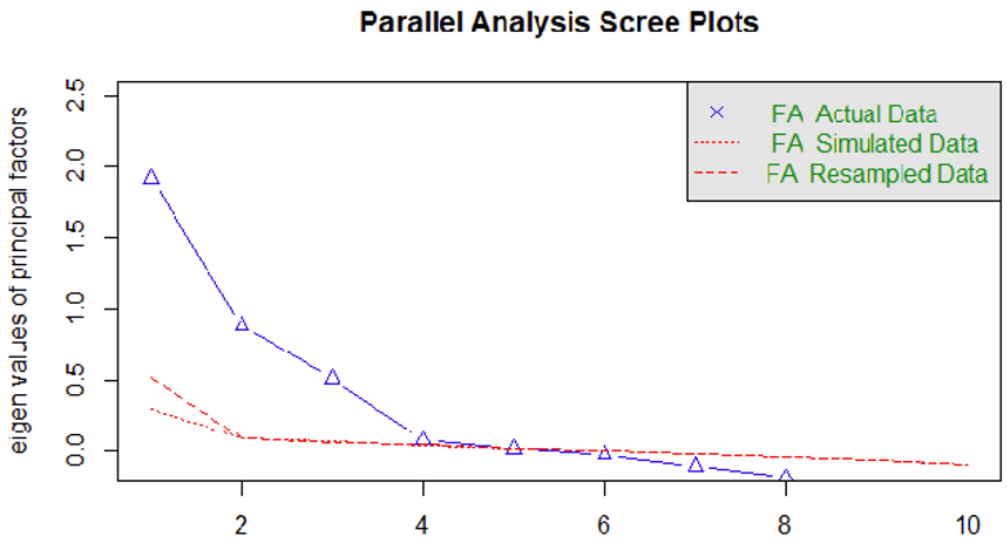



Figure 9: Parallel Analysis Scree Plots

Since the graph has an elbow point at 4, we decided to use four factors for the factor analysis.

Figure 10: Factor Analysis

```
Uniquenesses:
      Views       Likes Danceability      Energy    Loudness Speechiness    Liveness     Valence       Tempo
Duration_ms
      0.232       0.005       0.005       0.057       0.363       0.934       0.942       0.005       0.957
0.965

Loadings:
             Factor1 Factor2 Factor3 Factor4
Views         0.874
Likes         0.995
Danceability                  0.981   0.169
Energy                0.939   0.205   0.130
Loudness      0.102   0.714   0.336
Speechiness                   0.250
Liveness              0.222
Valence               0.232   0.324   0.915
Tempo                 0.160
Duration_ms                  -0.145

               Factor1 Factor2 Factor3 Factor4
SS loadings      1.775   1.532   1.324   0.904
Proportion Var   0.177   0.153   0.132   0.090
Cumulative Var   0.177   0.331   0.463   0.554

Test of the hypothesis that 4 factors are sufficient.
The chi square statistic is 63.66 on 11 degrees of freedom.
The p-value is 1.93e-09
```

Since the p-value is 1.93e-09 < 0.05, we reject the null hypothesis. Test of the hypothesis that four factors are sufficient. In the factor analysis, factors explain the variances of 17.7%, 15.3%, 13.2%, and 9%, respectively. Thus, a total of 55.4% of the variance is explained by these four factors.

### 3.6 Discrimination and Classification

```
Prior probabilities of groups:
      album compilation      single
 0.71847899  0.03002001  0.25150100

Group means:
                Views       Likes Danceability      Energy    Loudness  Speechiness    Liveness     Valence       Tempo
Duration_ms
album        0.01495253 -0.01545411  -0.05816903 -0.02076717 -0.03149407  0.001258569  0.02761236  0.01593482 -0.01879005
0.0780082
compilation -0.09388382 -0.15256878  -0.66932910 -0.44647630 -0.38791266 -0.322901822  0.12756231 -0.10475459 -0.24930214
0.2215202
single       0.03657148  0.13075570   0.26194840  0.10518826  0.18414164  0.043910294 -0.06266487 -0.05793308  0.06166321
-0.2717478

Coefficients of linear discriminants:
                    LD1         LD2
Views         -0.45904797  0.57983890
Likes          0.52011417 -0.61812081
Danceability   0.72304181  0.48096563
Energy         0.26704705  0.78035113
Loudness       0.20368943 -0.61267064
Speechiness   -0.06771667  0.37650952
Liveness      -0.16311476 -0.06352243
Valence       -0.63247891 -0.03250776
Tempo          0.23056036  0.21147177
Duration_ms   -0.52937215  0.51675571

Proportion of trace:
   LD1    LD2
0.9247 0.0753
```

Figure 12: Linear Discrimination Analysis for Album Type

The LDA output shows that 71.8% of the training observations correspond to albums, 3% compilations, and 25.1% singles. The accuracy of our model for train data is 72.1% and the accuracy of our model for test data is 68.1%.

### 3.7 Clustering

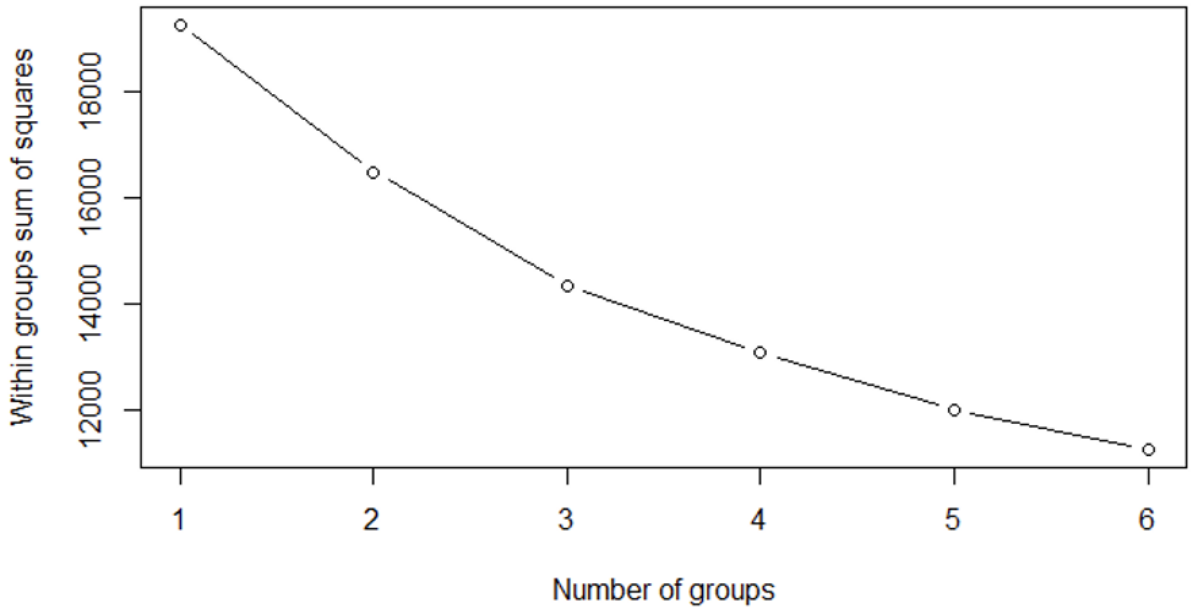


Figure 13: Plot for finding the best number of groups.

The elbow point is three (3).

| | Views | Likes | Danceability | Energy | Loudness | Speechiness | Liveness | Valence | Tempo | Duration_ms | album_type |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | -2.9785556 | -0.9599344 | -4.645878 | -12.0339173 | -13.796638 | -4.0374873 | -0.6609207 | -1.911684 | -1.3786009 | 2.6069247 | 9.501968 |
| 2 | 72.2625111 | 47.0070265 | 1.644932 | 0.6067044 | 3.074851 | 2.1376252 | -0.8147114 | 1.841044 | 1.2640953 | 0.2235199 | 14.626280 |
| 3 | -0.5222029 | -0.8003287 | 2.195466 | 6.5590277 | 1.515723 | 0.8626097 | 0.7214734 | 5.267968 | 0.7142432 | -0.4331101 | 2.648204 |

Figure 14: K-means clustering with three groups.

```
 [1] 0.6135046 0.6158200 0.6158209 0.6158210 0.6158210 0.6158210 0.6158464 0.6167815 0.6198726
[10] 1.0000000
```

Figure 15: Vector of cumulative proportions of variance

### 3.8 Canonical Correlation Analysis

```
Wilks' Lambda, using F-approximation (Rao's F):
              stat   approx df1  df2      p.value
1 to 2:  0.9500592 5.376671  18 3730 1.320721e-12
2 to 2:  0.9832177 3.981285   8 1866 1.083789e-04
            [,1]      [,2]
Views  0.684202 -1.931863
Likes -1.539853  1.352435
                      [,1]        [,2]
Danceability -0.4314270201 -0.30079985
Energy        0.0914661029 -0.11105923
Loudness     -0.8096211697 -0.01434253
Speechiness  -0.1889310072  0.41755705
Liveness      0.1572095325  0.09686818
Valence       0.5457077423 -0.53746855
Tempo        -0.1937711883 -0.26164532
Duration_ms  -0.0006031122 -0.35756607
album_type   -0.3831680444  0.68863231
```

Figure 16: Canonical correlation analysis

P-values are significant for both dimensions.

## 4. Discussion/Conclusion

The analysis revealed nuanced insights into digital music consumption, emphasizing the impact of album types on danceability and the interconnected influences of energy, liveness, and valence. Statistical tests bolstered result robustness, though caution is advised, considering potential deviations from normality. The study provides valuable considerations for music promotion and platform algorithms, with the potential for deeper exploration into statistical

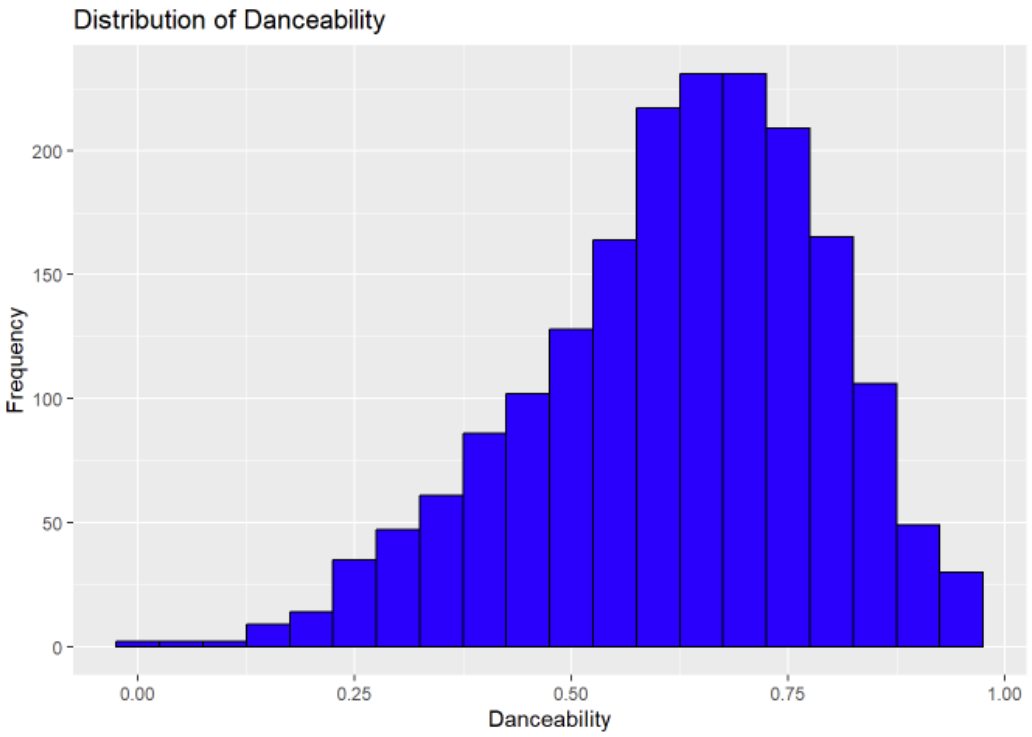
nuances and a broader examination of user behavior given more time. Aligning with existing literature, this research contributes to understanding digital music consumption dynamics, laying the groundwork for future investigations.

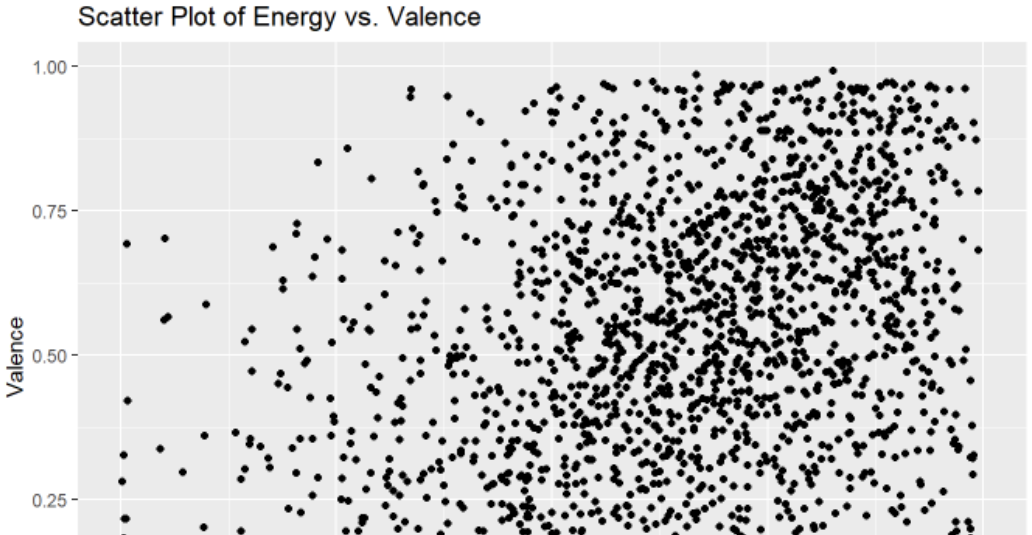
## References

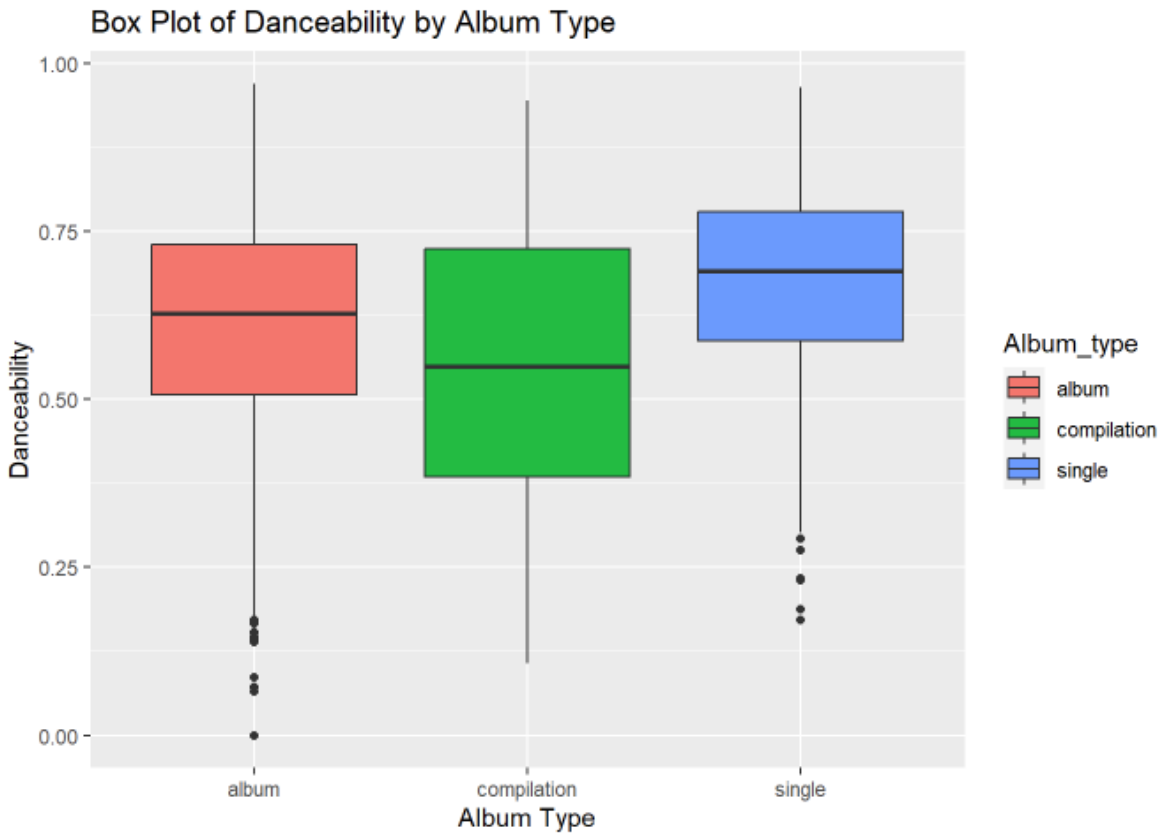

Rastelli, S. (2023). Spotify and YouTube. Kaggle. https://www.kaggle.com/datasets/salvatorerastelli/spotify-and-youtube

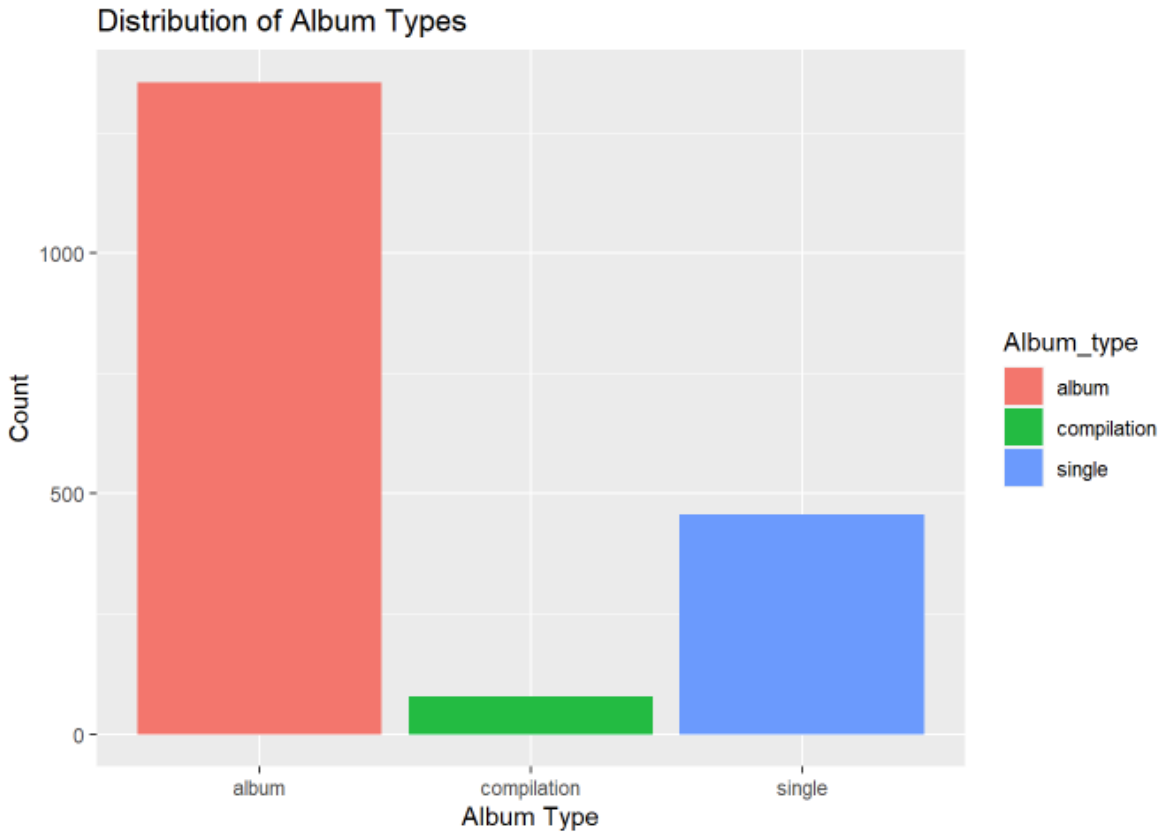
UCLA Statistical Consulting Group. (Accessed December 30, 2023). Canonical Correlation Analysis. Retrieved from https://stats.oarc.ucla.edu/r/dae/canonical-correlation-analysis/

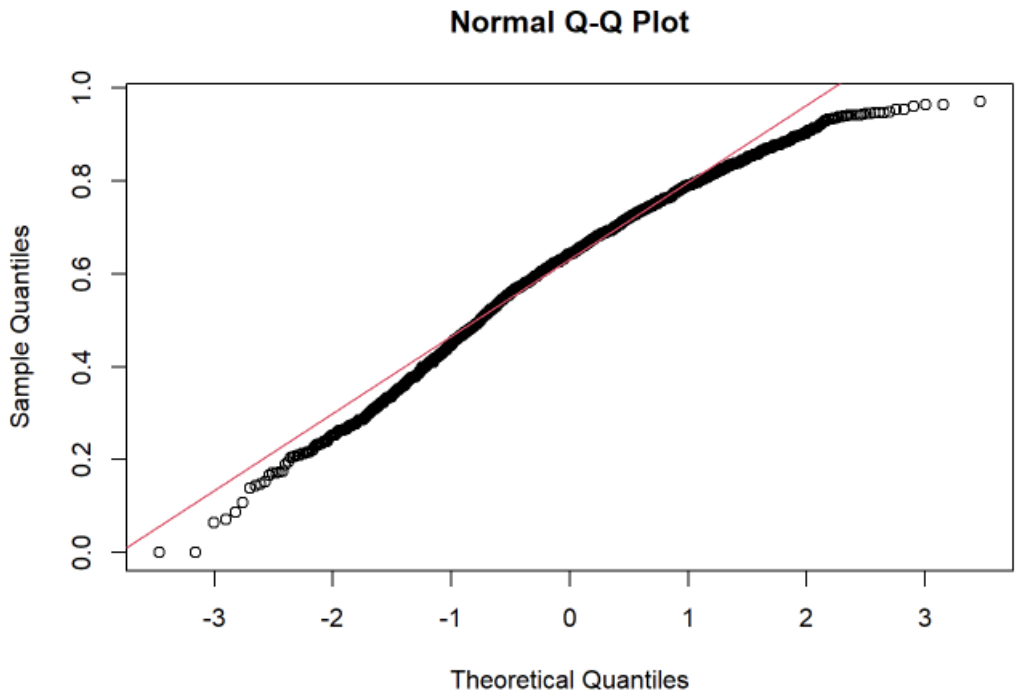

## Appendices



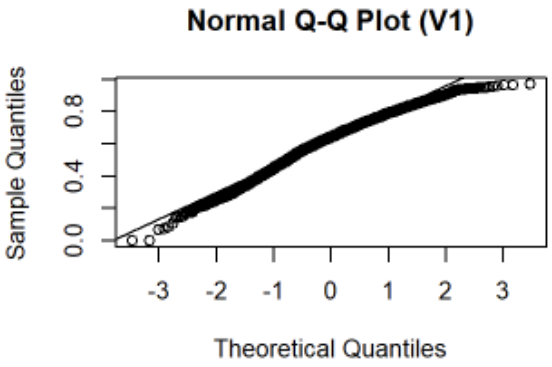



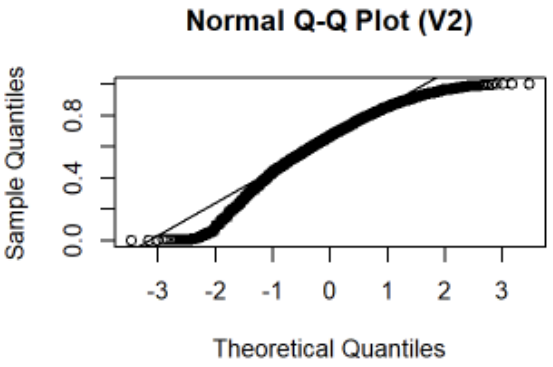

Box Plot of Danceability by Album Type
1.00
0.75
0.50
0.25
0.00
Danceability
album
compilation
single
Album Type
Album_type
album
compilation
single

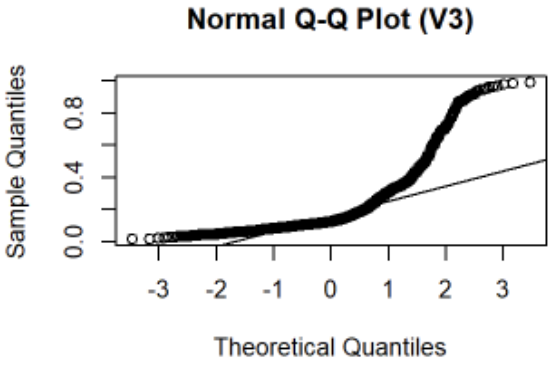

Distribution of Album Types
1000
500
0
Count
album
compilation
single
Album Type
Album_type
album
compilation
single

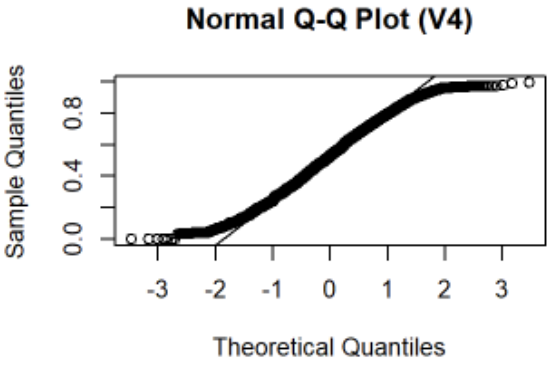


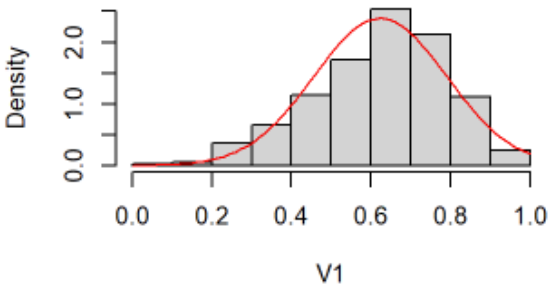

```
##      Test        H p value MVN
## 1 Royston 2541.242       0  NO
```

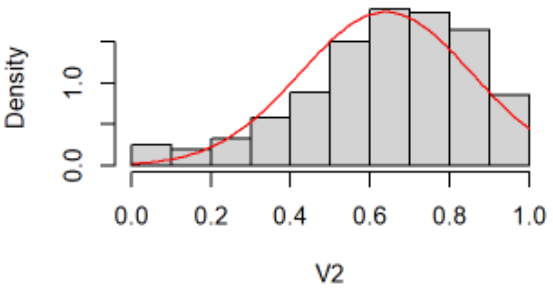

Normal Q-Q Plot (V1)
Sample Quantiles
0.0
0.4
0.8
-3
-2
-1
0
1
2
3
Theoretical Quantiles

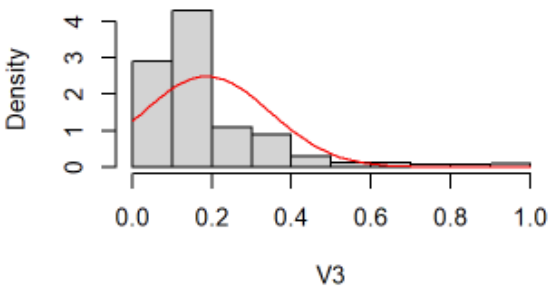

Normal Q-Q Plot (V2)
Sample Quantiles
0.0
0.4
0.8
-3
-2
-1
0
1
2
3
Theoretical Quantiles

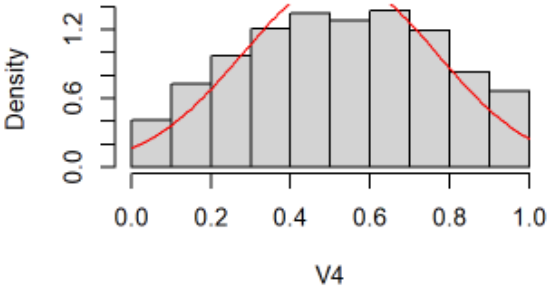

Normal Q-Q Plot (V3)
Sample Quantiles
0.0
0.4
0.8
-3
-2
-1
0
1
2
3
Theoretical Quantiles

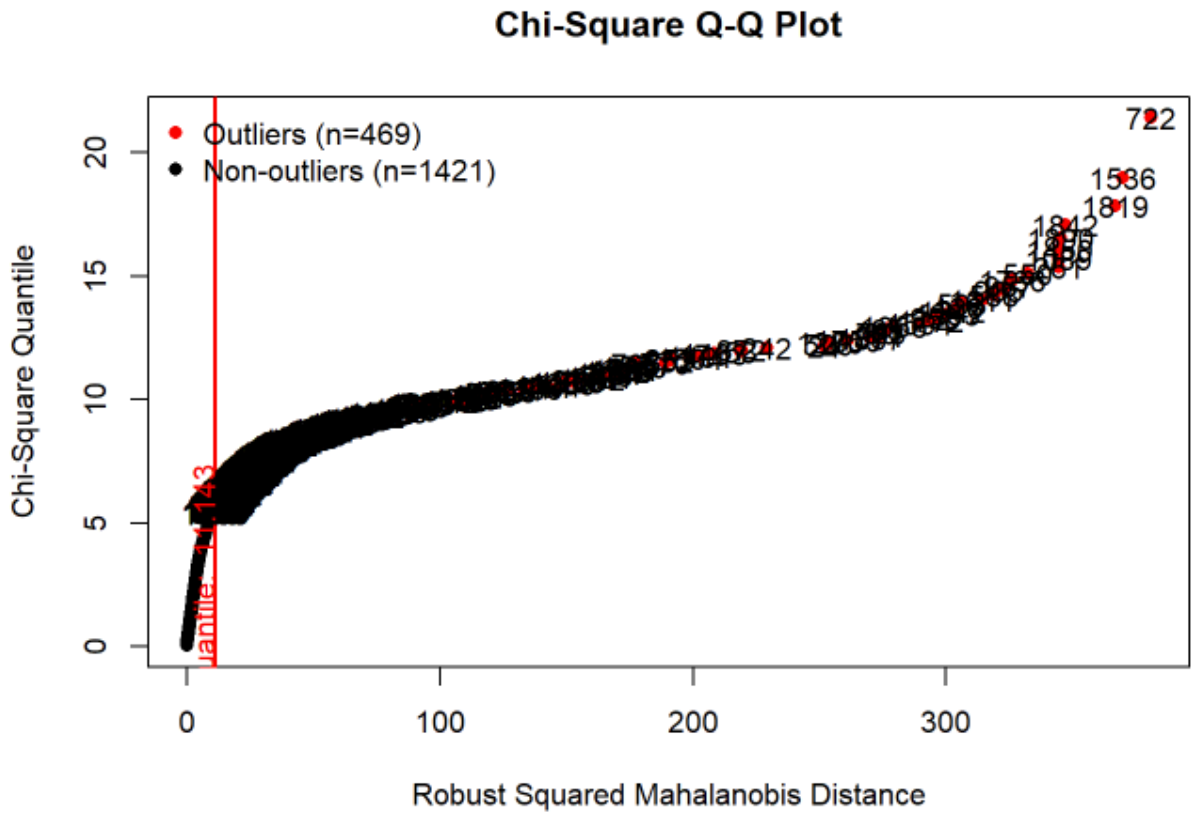

Normal Q-Q Plot (V4)
Sample Quantiles
0.0
0.4
0.8
-3
-2
-1
0
1
2
3
Theoretical Quantiles


Density
0.0
1.0
2.0
0.0
0.2
0.4
0.6
0.8
1.0
V1


Density
0.0
1.0
0.0
0.2
0.4
0.6
0.8
1.0
V2


Density
0
1
2
3
4
0.0
0.2
0.4
0.6
0.8
1.0
V3


Density
0.0
0.6
1.2
0.0
0.2
0.4
0.6
0.8
1.0
V4




```
## Danceability      Valence
##    0.6226546    0.5234526
```

```
##              Test         Statistic              p value Result
## 1 Mardia Skewness  116.322719648457 3.25747068499248e-24     NO
## 2 Mardia Kurtosis -4.18047096962961 2.90905950337539e-05     NO
## 3             MVN              <NA>                  <NA>     NO
```

```
##              Test     Variable Statistic   p value Normality
## 1 Anderson-Darling Danceability   51.7933  <0.001      NO
## 2 Anderson-Darling   Valence      82.0245  <0.001      NO
```

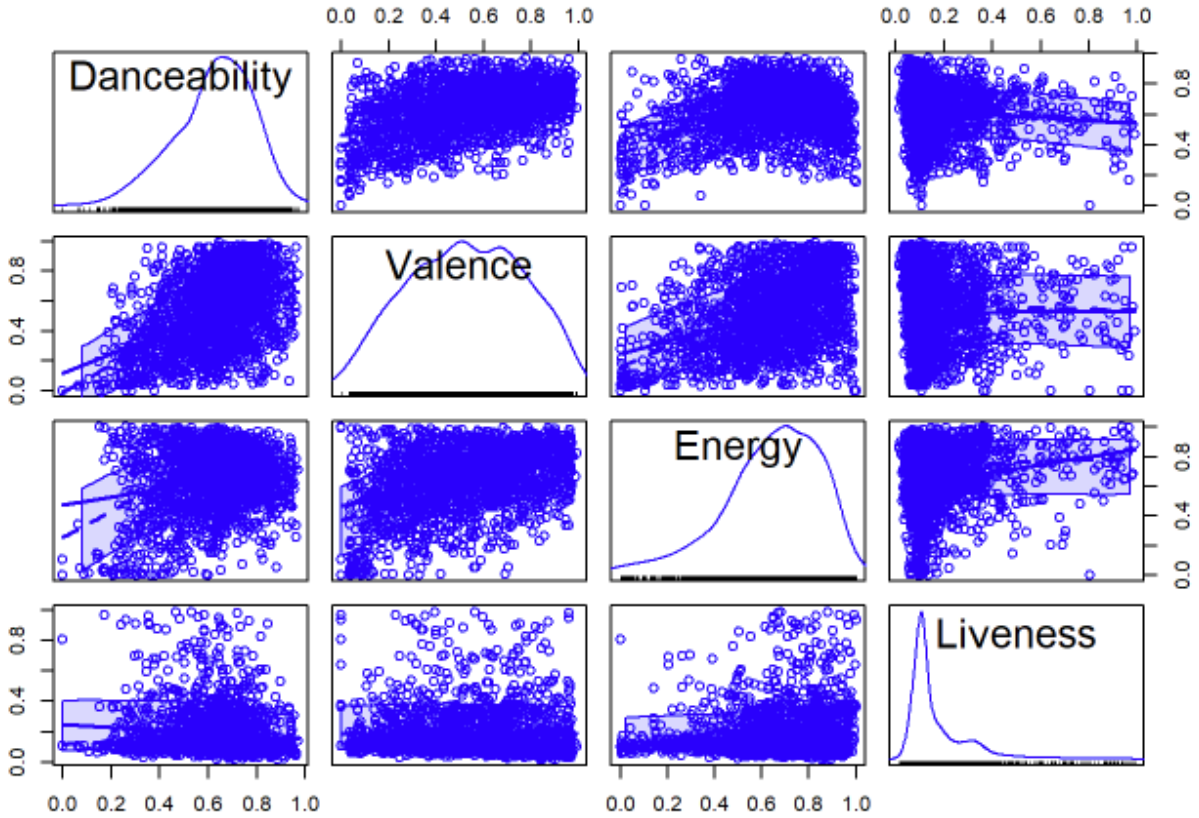
```
##
##  Hotelling's one sample T2-test
##
## data:  y
## T.2 = 514.44, df1 = 2, df2 = 1888, p-value < 2.2e-16
## alternative hypothesis: true location is not equal to c(0.7,0.7)
```

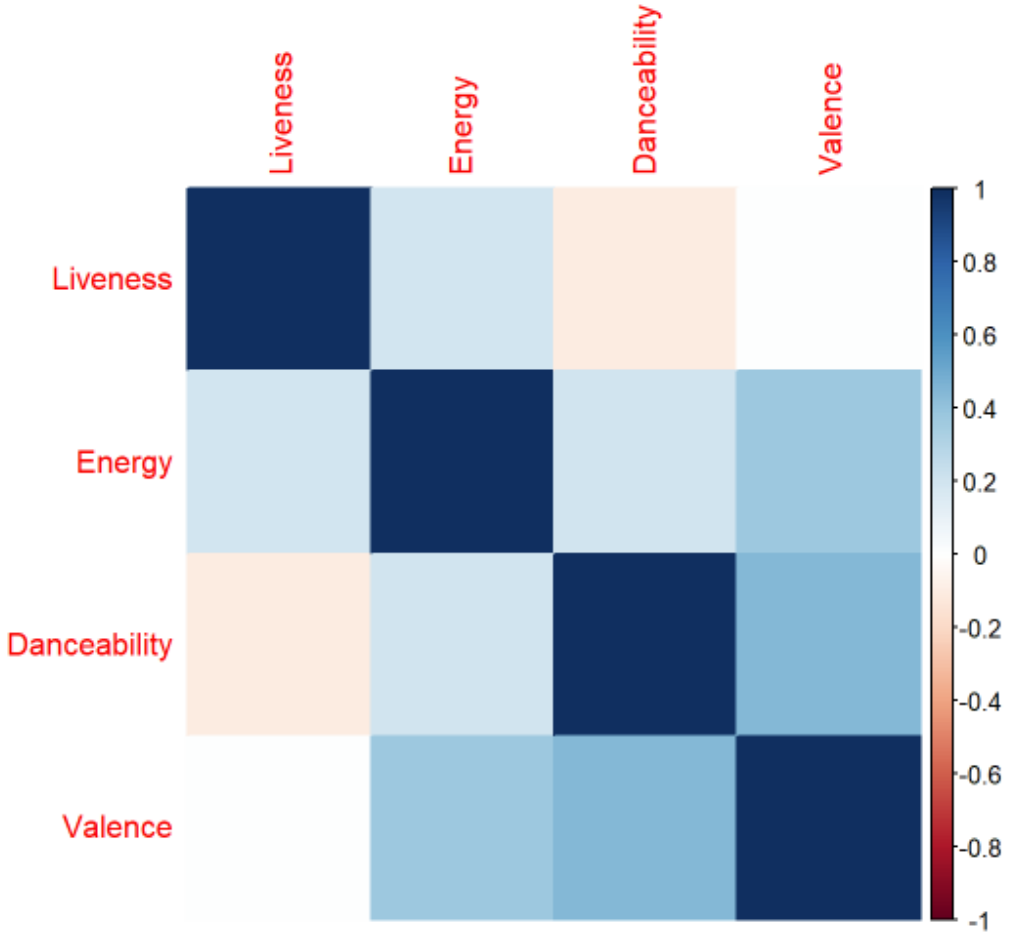
```
##              Df   Pillai approx F num Df den Df    Pr(>F)
## Album_type    2 0.059218   14.379      8   3770 < 2.2e-16 ***
## Residuals  1887
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##  Response Danceability :
##               Df Sum Sq Mean Sq F value   Pr(>F)
## Album_type     2  1.587 0.79340   29.24 3.12e-13 ***
## Residuals   1887 51.202 0.02713
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##  Response Energy :
##               Df Sum Sq Mean Sq F value    Pr(>F)
## Album_type     2  1.329 0.66467   14.69 4.669e-07 ***
## Residuals   1887 85.378 0.04525
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##  Response Liveness :
##               Df Sum Sq  Mean Sq F value Pr(>F)
## Album_type     2  0.070 0.034873   1.364 0.2559
## Residuals   1887 48.243 0.025566
##
##  Response Valence :
##               Df  Sum Sq  Mean Sq F value Pr(>F)
## Album_type     2   0.107 0.053283  0.8888 0.4113
## Residuals   1887 113.119 0.059946
```

```
## 'data.frame':    1890 obs. of  14 variables:
##  $ X           : int  18846 18894 2985 1841 3370 11637 6745 16127 2756 12635 ...
##  $ Danceability: num  0.896 0.786 0.631 0.669 0.737 0.751 0.598 0.562 0.323 0.784 ...
##  $ Energy      : num  0.818 0.572 0.631 0.829 0.603 0.746 0.824 0.46 0.0114 0.73 ...
##  $ Key         : num  11 11 6 1 10 2 11 5 4 1 ...
##  $ Loudness    : num  -5.67 -4.97 -6.93 -3.8 -7.49 ...
##  $ Speechiness : num  0.191 0.0399 0.0307 0.49 0.239 0.292 0.131 0.259 0.0343 0.191 ...
##  $ Liveness    : num  0.101 0.0808 0.133 0.241 0.118 0.0503 0.0975 0.107 0.171 0.107 ...
##  $ Valence     : num  0.804 0.539 0.697 0.61 0.492 0.889 0.416 0.623 0.134 0.0415 ...
##  $ Tempo       : num  90 92 90 119 76 ...
##  $ Duration_ms : num  205333 212800 263152 253067 281800 ...
##  $ Views       : num  313181 4232418 87878430 5174815 48682 ...
##  $ Likes       : num  3962 179601 478253 67542 461 ...
##  $ Comments    : num  90 2907 12162 3536 20 ...
##  $ Stream      : num  6.95e+05 2.35e+07 4.17e+07 1.02e+08 6.17e+04 ...
##  - attr(*, "na.action")= 'omit' Named int [1:110] 7 30 32 48 57 97 103 107 122 137 ...
##   ..- attr(*, "names")= chr [1:110] "4761" "3004" "7989" "14426" ...
```

```
## [1] 1890   14
```


Danceability
Valence
Energy
Liveness
0.0 0.2 0.4 0.6 0.8 1.0

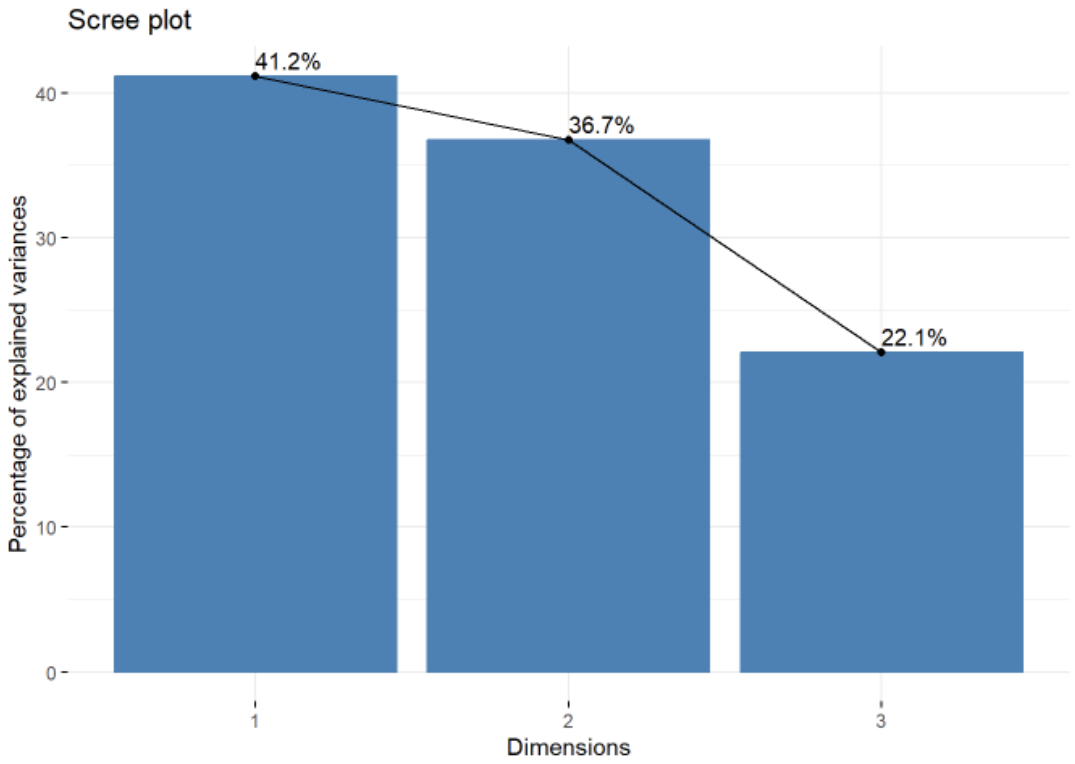

Liveness
Energy
Danceability
Valence
1
0.8
0.6
0.4
0.2
0
-0.2
-0.4
-0.6
-0.8
-1

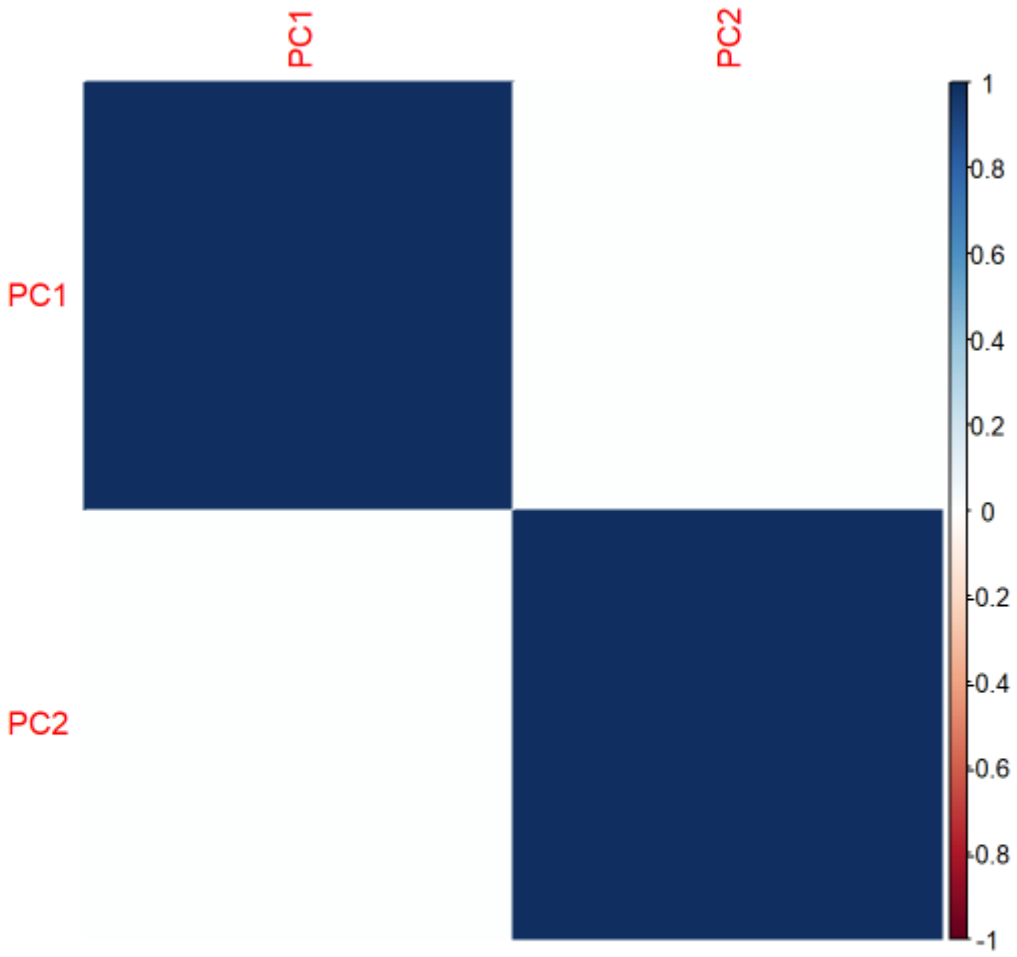
```
## attr(,"scaled:scale")
## Danceability      Valence       Energy     Liveness
##    0.1671694    0.2448251    0.2142456    0.1599240
```

```
##              Danceability     Valence    Energy     Liveness
## Danceability    1.0000000 0.446799508 0.2031253 -0.102530536
## Valence         0.4467995 1.000000000 0.3794897  0.008407797
## Energy          0.2031253 0.379489675 1.0000000  0.195936879
## Liveness       -0.1025305 0.008407797 0.1959369  1.000000000
```

```
## Importance of components:
##                            PC1    PC2    PC3
## Standard deviation     1.1116 1.0500 0.8136
## Proportion of Variance 0.4119 0.3675 0.2206
## Cumulative Proportion  0.4119 0.7794 1.0000
```

```
##                    PC1         PC2        PC3
## Danceability 0.4735516 -0.68949253  0.5480410
## Energy       0.7674845  0.01776261 -0.6408214
## Liveness     0.4321069  0.72407495  0.5375864
```

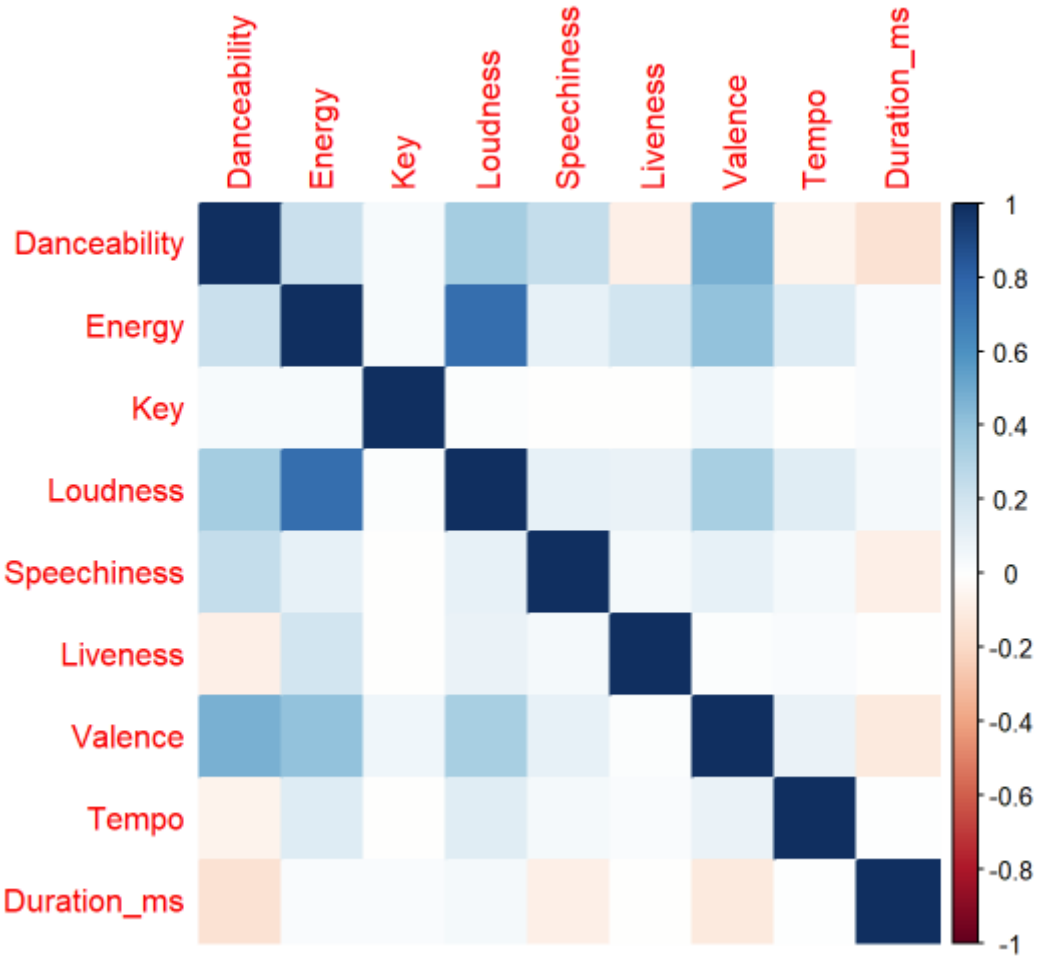


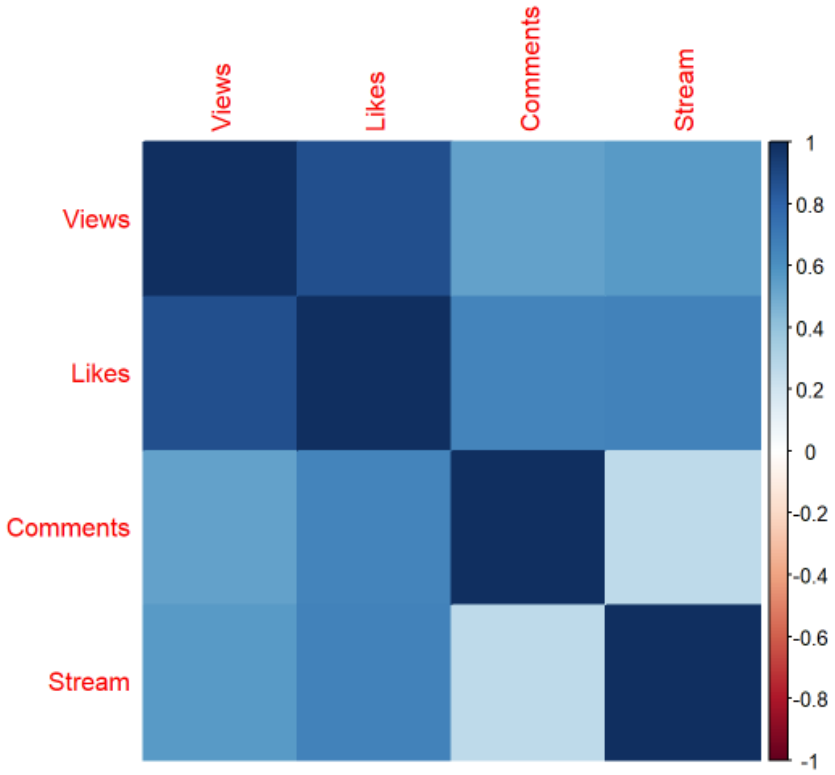



```
##                      PC1          PC2
## Danceability 0.5263985 -0.72394753
## Valence      0.4556221 -0.28118508
## Energy       0.8531334  0.01865023
## Liveness     0.4803287  0.76025809
```

```
##
## Call:
## lm(formula = base_total ~ ., data = ols.data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.62190 -0.59256 -0.00188  0.61268  2.37122
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.148e-16  1.944e-02    0.00        1
## PC1          4.099e-01  1.749e-02   23.43   <2e-16 ***
## PC2         -2.678e-01  1.852e-02  -14.46   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.845 on 1887 degrees of freedom
## Multiple R-squared:  0.2867, Adjusted R-squared:  0.2859
## F-statistic: 379.1 on 2 and 1887 DF,  p-value: < 2.2e-16
```

```
##        X            Artist          Url_spotify          Track
##  Min.   :    3   Length:1876        Length:1876        Length:1876
##  1st Qu.: 5328   Class :character   Class :character   Class :character
##  Median :10362   Mode  :character   Mode  :character   Mode  :character
##  Mean   :10504
##  3rd Qu.:15924
##  Max.   :20697
##     Album            Album_type           Uri             Danceability
##  Length:1876        Length:1876        Length:1876        Min.   :0.0000
##  Class :character   Class :character   Class :character   1st Qu.:0.5180
##  Mode  :character   Mode  :character   Mode  :character   Median :0.6395
##                                                           Mean   :0.6192
##                                                           3rd Qu.:0.7402
##                                                           Max.   :0.9700
##      Energy              Key            Loudness        Speechiness
##  Min.   :0.000055   Min.   : 0.000   Min.   :-44.761   Min.   :0.00000
##  1st Qu.:0.514500   1st Qu.: 2.000   1st Qu.: -8.755   1st Qu.:0.03508
##  Median :0.673000   Median : 6.000   Median : -6.496   Median :0.05100
##  Mean   :0.639190   Mean   : 5.336   Mean   : -7.692   Mean   :0.09622
##  3rd Qu.:0.795250   3rd Qu.: 8.000   3rd Qu.: -4.929   3rd Qu.:0.10225
##  Max.   :1.000000   Max.   :11.000   Max.   : -0.514   Max.   :0.94700
##     Liveness          Valence           Tempo          Duration_ms
##  Min.   :0.01900   Min.   :0.0000   Min.   :  0.00   Min.   : 31000
##  1st Qu.:0.09595   1st Qu.:0.3440   1st Qu.: 97.03   1st Qu.:180939
##  Median :0.12900   Median :0.5470   Median :118.98   Median :213063
##  Mean   :0.19958   Mean   :0.5350   Mean   :119.83   Mean   :222418
##  3rd Qu.:0.24325   3rd Qu.:0.7402   3rd Qu.:137.99   3rd Qu.:249236
##  Max.   :0.99000   Max.   :0.9840   Max.   :213.50   Max.   :824133
##  Url_youtube          Title              Channel             Views
##  Length:1876        Length:1876        Length:1876        Min.   :1.380e+02
##  Class :character   Class :character   Class :character   1st Qu.:2.134e+06
##  Mode  :character   Mode  :character   Mode  :character   Median :1.689e+07
##                                                           Mean   :9.928e+07
##                                                           3rd Qu.:8.208e+07
##                                                           Max.   :4.821e+09
##      Likes             Comments         Licensed         official_video
##  Min.   :       1   Min.   :      0   Length:1876        Length:1876
##  1st Qu.:   24256   1st Qu.:    558   Class :character   Class :character
##  Median :  146207   Median :   3760   Mode  :character   Mode  :character
##  Mean   :  695949   Mean   :  30054
##  3rd Qu.:  594353   3rd Qu.:  16380
##  Max.   :26399133   Max.   :5331537
##      Stream
##  Min.   :7.771e+03
##  1st Qu.:1.852e+07
##  Median :4.999e+07
##  Mean   :1.351e+08
##  3rd Qu.:1.377e+08
##  Max.   :2.595e+09
```


Danceability
Energy
Key
Loudness
Speechiness
Liveness
Valence
Tempo
Duration_ms
1
0.8
0.6
0.4
0.2
0
-0.2
-0.4
-0.6
-0.8
-1


Views
Likes
Comments
Stream
1
0.8
0.6
0.4
0.2
0
-0.2
-0.4
-0.6
-0.8
-1

```
##           Views     Likes
## Views 1.0000000 0.8728789
## Likes 0.8728789 1.0000000
```

```
##
## Factor analysis with Call: fa(r = spotify_scaled[, c("Views", "Likes")], nfactors = 1, rot
ate = "varimax")
##
## Test of the hypothesis that 1 factor is sufficient.
## The degrees of freedom for the model is -1  and the objective function was  0
## The number of observations was  1876  with Chi Square =  0  with prob <  NA
##
## The root mean square of the residuals (RMSA) is  0
## The df corrected root mean square of the residuals is  NA
##
## Tucker Lewis Index of factoring reliability =  1
```

```
##
## Loadings:
##       MR1
## Views 0.934
## Likes 0.934
##
##                  MR1
## SS loadings    1.746
## Proportion Var 0.873
```

```
##
## Factor analysis with Call: fa(r = spotify_scaled[, c("Views", "Likes")], nfactors = 1, rot
ate = "promax")
##
## Test of the hypothesis that 1 factor is sufficient.
## The degrees of freedom for the model is -1  and the objective function was  0
## The number of observations was  1876  with Chi Square =  0  with prob <  NA
##
## The root mean square of the residuals (RMSA) is  0
## The df corrected root mean square of the residuals is  NA
##
## Tucker Lewis Index of factoring reliability =  1
```

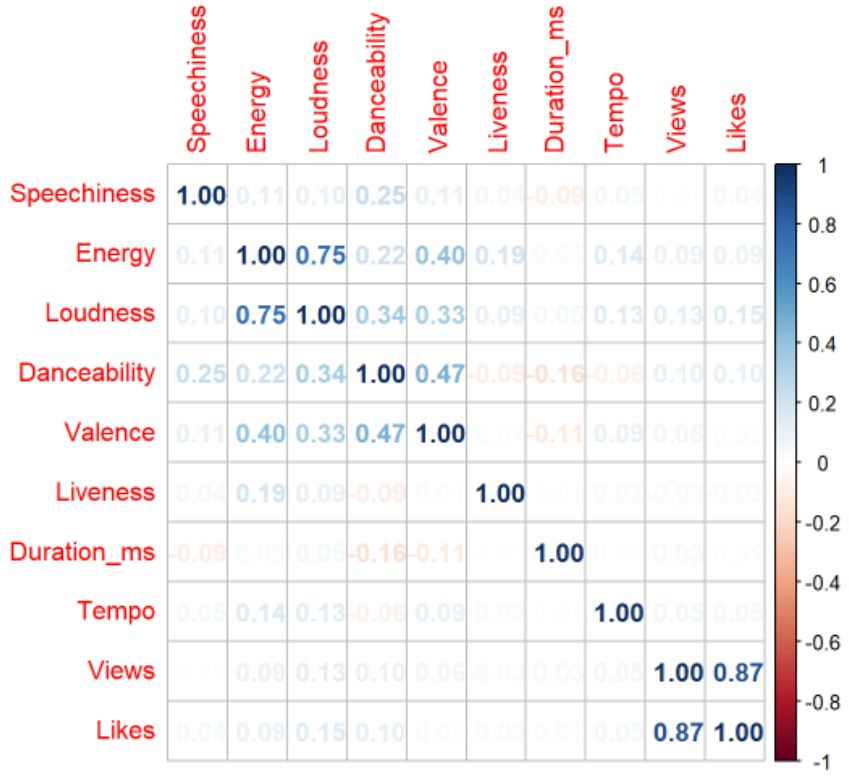
```
##
## Loadings:
##       MR1
## Views 0.934
## Likes 0.934
##
##                  MR1
## SS loadings    1.746
## Proportion Var 0.873
```

```
##
## Call:
## factanal(x = spotify_scaled, factors = 1, rotation = "varimax")
##
## Uniquenesses:
##        Views        Likes Danceability       Energy     Loudness  Speechiness
##        0.144        0.111        0.987        0.989        0.975        0.999
##     Liveness      Valence        Tempo  Duration_ms
##        0.999        0.998        0.997        1.000
##
## Loadings:
##              Factor1
## Views         0.925
## Likes         0.943
## Danceability  0.115
## Energy        0.106
## Loudness      0.157
## Speechiness
## Liveness
## Valence
## Tempo
## Duration_ms
##
##                Factor1
## SS loadings      1.802
## Proportion Var   0.180
##
## Test of the hypothesis that 1 factor is sufficient.
## The chi square statistic is 2946.44 on 35 degrees of freedom.
## The p-value is 0
```

```
## subset_res_c
##    1    2
## 1833   43
```

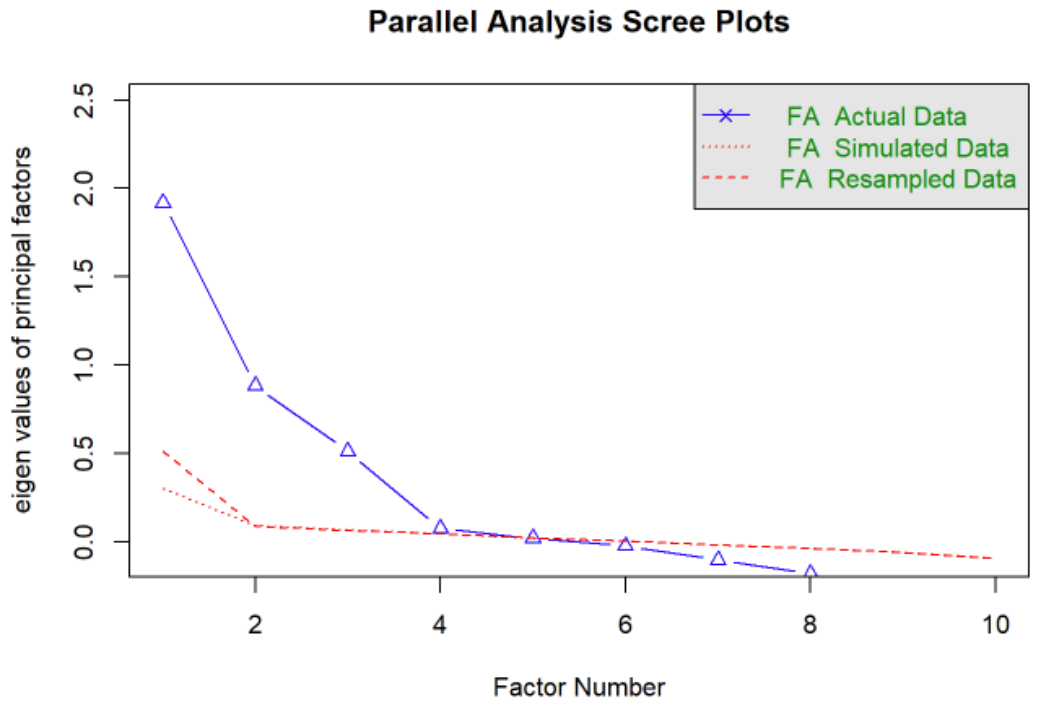
```
##         Length Class  Mode
## prior    2     -none- numeric
## counts   2     -none- numeric
## means   16     -none- numeric
## scaling  8     -none- numeric
## lev      2     -none- character
## svd      1     -none- numeric
## N        1     -none- numeric
## call     3     -none- call
## terms    3     terms  call
## xlevels  0     -none- list
```

```
## [1] 1876   10
```



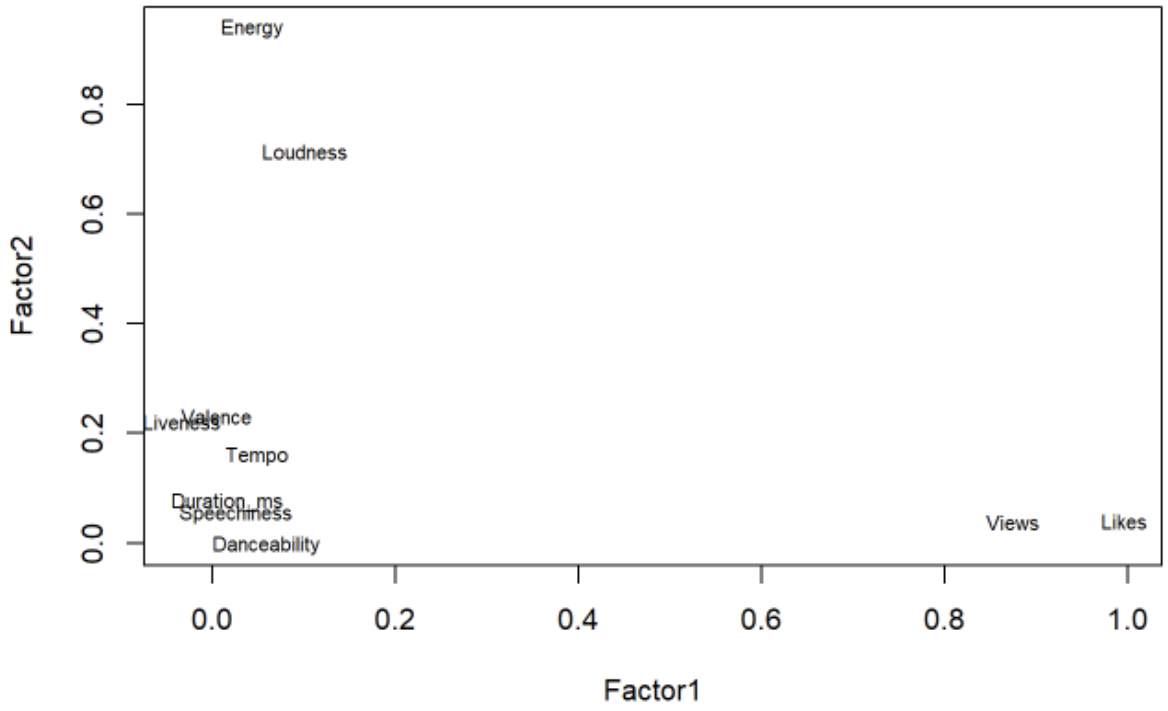

```
## Kaiser-Meyer-Olkin factor adequacy
## Call: KMO(r = cm)
## Overall MSA =  0.56
## MSA for each item =
##        Views        Likes Danceability       Energy     Loudness  Speechiness
##         0.51         0.51         0.57         0.57         0.59         0.60
##     Liveness      Valence        Tempo  Duration_ms
##         0.54         0.64         0.56         0.58
```

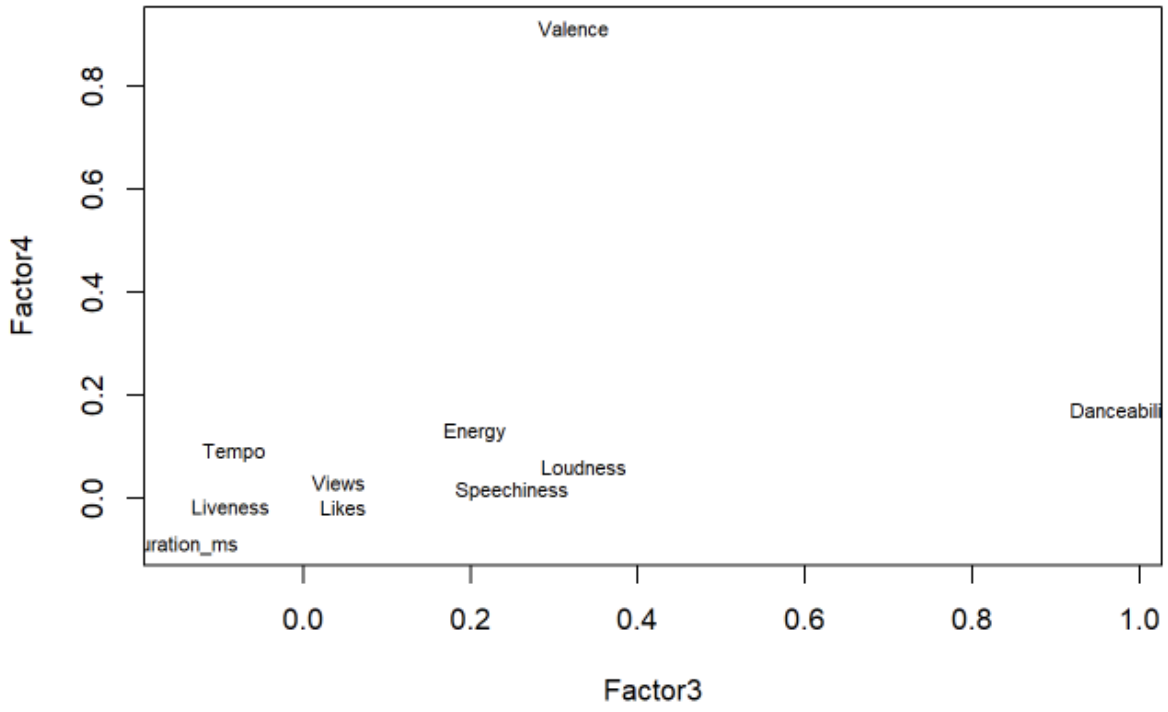
```
## $chisq
## [1] 5724.413
##
## $p.value
## [1] 0
##
## $df
## [1] 45
```




```
## Parallel analysis suggests that the number of factors =  4  and the number of components =
NA
```

```
##
## Call:
## factanal(x = spotify_scaled, factors = 4)
##
## Uniquenesses:
##        Views        Likes Danceability       Energy     Loudness  Speechiness
##        0.232        0.005        0.005        0.057        0.363        0.934
##     Liveness      Valence        Tempo  Duration_ms
##        0.942        0.005        0.957        0.965
##
## Loadings:
##              Factor1 Factor2 Factor3 Factor4
## Views         0.874
## Likes         0.995
## Danceability                  0.981   0.169
## Energy                0.939   0.205   0.130
## Loudness      0.102   0.714   0.336
## Speechiness                   0.250
## Liveness              0.222
## Valence               0.232   0.324   0.915
## Tempo                 0.160
## Duration_ms                  -0.145
##
##                Factor1 Factor2 Factor3 Factor4
## SS loadings      1.775   1.532   1.324   0.904
## Proportion Var   0.177   0.153   0.132   0.090
## Cumulative Var   0.177   0.331   0.463   0.554
##
## Test of the hypothesis that 4 factors are sufficient.
## The chi square statistic is 63.66 on 11 degrees of freedom.
## The p-value is 1.93e-09
```


Energy
Loudness
Valence
Tempo
Duration_ms
Speechiness
Danceability
Views
Likes
Factor2
Factor1
0.0
0.2
0.4
0.6
0.8
1.0


Valence
Energy
Tempo
Loudness
Views
Speechiness
Likes
Liveness
Factor4
Factor3
0.0
0.2
0.4
0.6
0.8
1.0

```
## [1] "Views" "Likes"
```

```
## [1] "Energy"   "Loudness"
```

```
## [1] "Danceability"
```

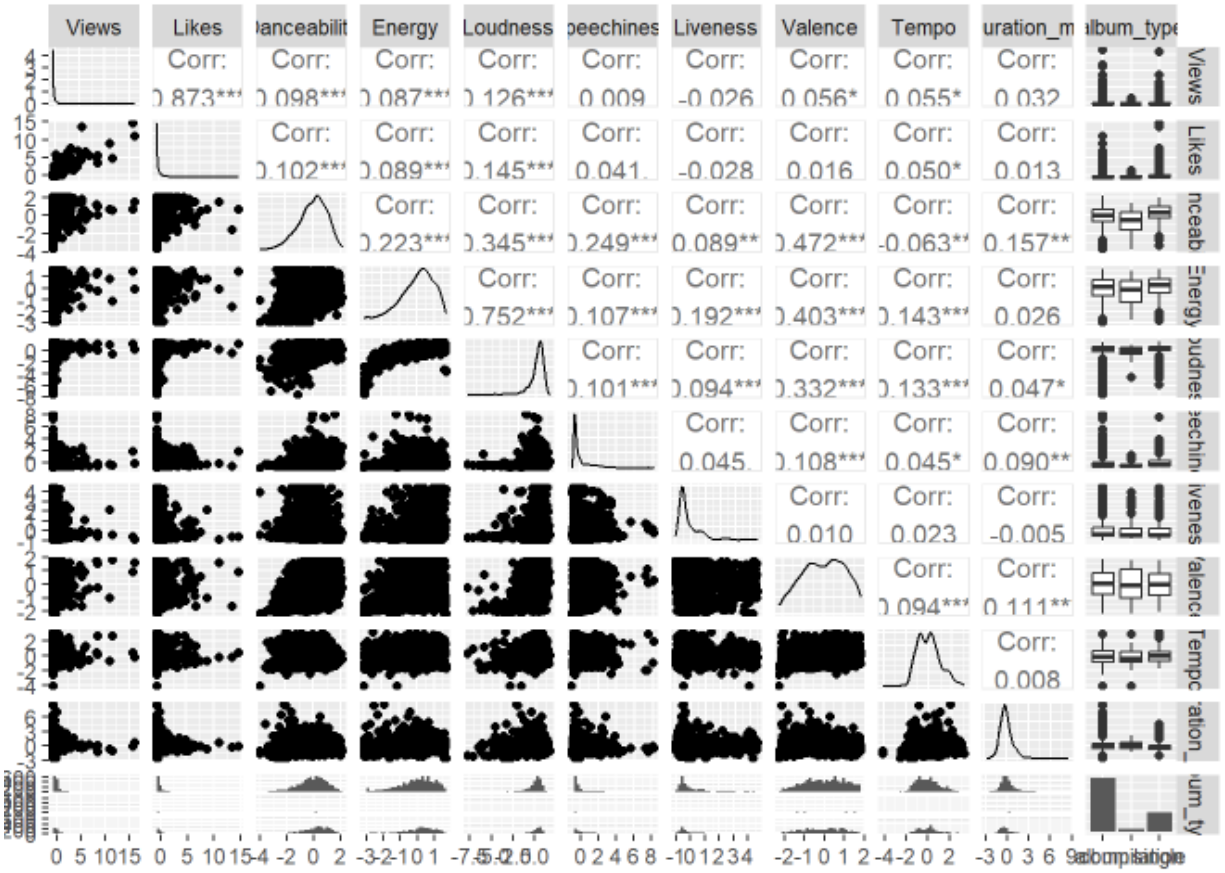
```
## [1] "Valence"
```

```
## Number of categories should be increased  in order to count frequencies.
```

```
##
## Reliability analysis
##  raw_alpha std.alpha G6(smc) average_r S/N    ase     mean   sd median_r
##       0.93      0.93    0.87      0.87  14 0.0031 -8.9e-18 0.97     0.87
```

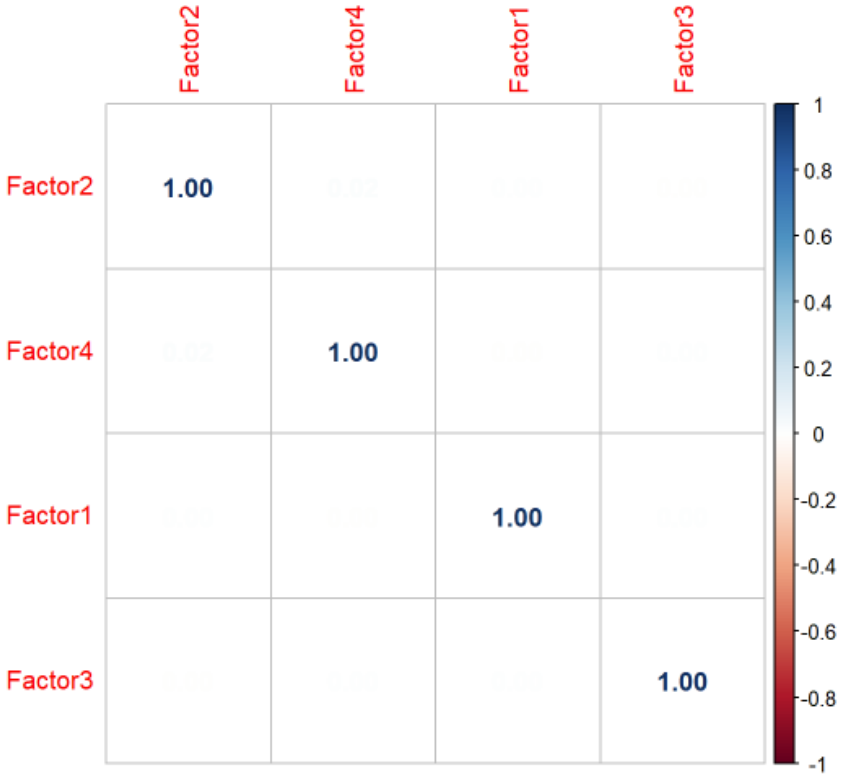
```
## Number of categories should be increased  in order to count frequencies.
```

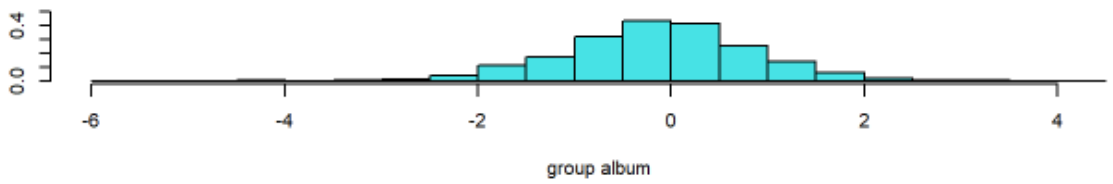
```
##
## Reliability analysis
##  raw_alpha std.alpha G6(smc) average_r S/N    ase     mean   sd median_r
##       0.86      0.86    0.75      0.75 6.1 0.0065 -1.8e-17 0.94     0.75
```

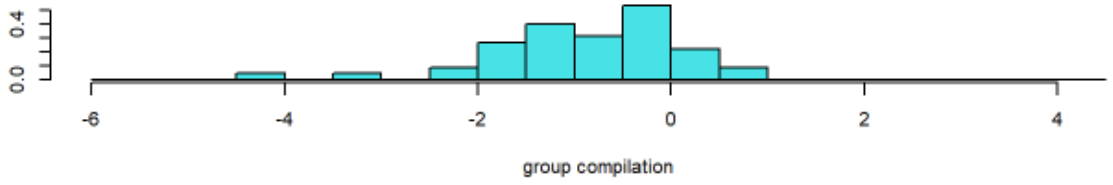


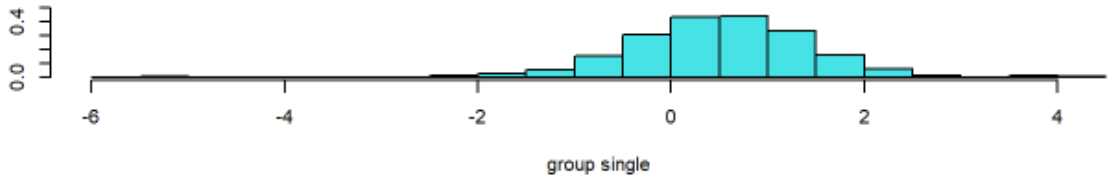
```
## Call:
## lda(album_type ~ ., data = train)
##
## Prior probabilities of groups:
##       album compilation      single
##  0.71847899  0.03002001  0.25150100
##
## Group means:
##                    Views       Likes Danceability      Energy    Loudness
## album        0.01495253 -0.01545411  -0.05816903 -0.02076717 -0.03149407
## compilation -0.09388382 -0.15256878  -0.66932910 -0.44647630 -0.38791266
## single       0.03657148  0.13075570   0.26194840  0.10518826  0.18414164
##              Speechiness    Liveness     Valence       Tempo Duration_ms
## album        0.001258569  0.02761236  0.01593482 -0.01879005   0.0780082
## compilation -0.322901822  0.12756231 -0.10475459 -0.24930214   0.2215202
## single       0.043910294 -0.06266487 -0.05793308  0.06166321  -0.2717478
##
## Coefficients of linear discriminants:
##                       LD1         LD2
## Views        -0.45904797  0.57983890
## Likes         0.52011417 -0.61812081
## Danceability  0.72304181  0.48096563
## Energy        0.26704705  0.78035113
## Loudness      0.20368943 -0.61267064
## Speechiness  -0.06771667  0.37650952
## Liveness     -0.16311476 -0.06352243
## Valence      -0.63247891 -0.03250776
## Tempo         0.23056036  0.21147177
## Duration_ms  -0.52937215  0.51675571
##
## Proportion of trace:
##    LD1    LD2
## 0.9247 0.0753
```

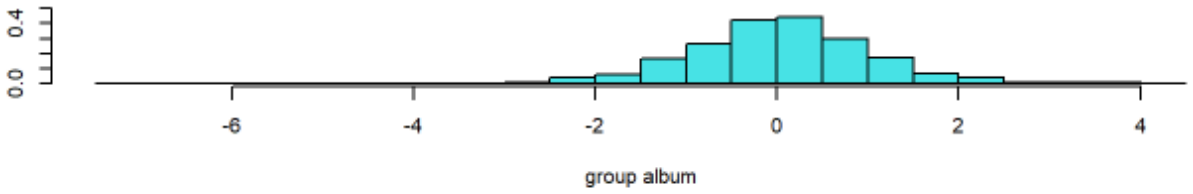
```
##          Factor1     Factor2     Factor3     Factor4
## 3285  -0.2191333 -1.30014227 -2.84141636 -0.75020731
## 8041   1.1873374 -1.13493514  1.61742805 -1.02439238
## 16463 -0.1849736  0.08479401 -1.06407429 -0.53797823
## 6934  -0.3313377 -0.88129281  1.21616384  1.70689445
## 4476  -0.2907244 -1.14146475 -0.02142963 -0.05474933
## 14983 -0.1610969 -0.98744398 -2.00750135 -0.73073557
```

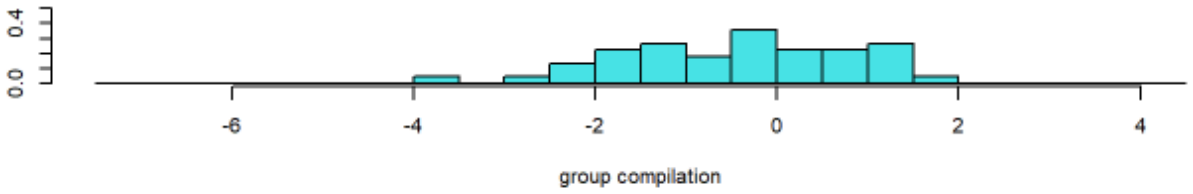


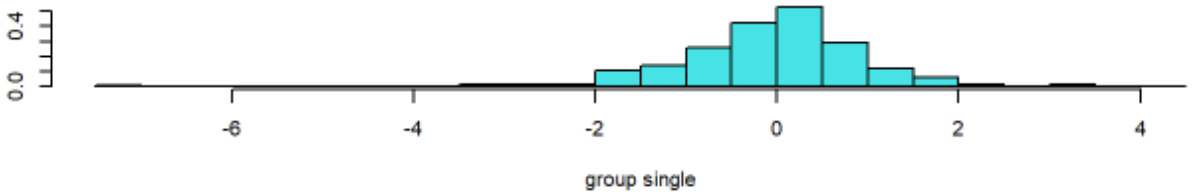

0.0
0.4
-6
-4
-2
0
2
4
group album

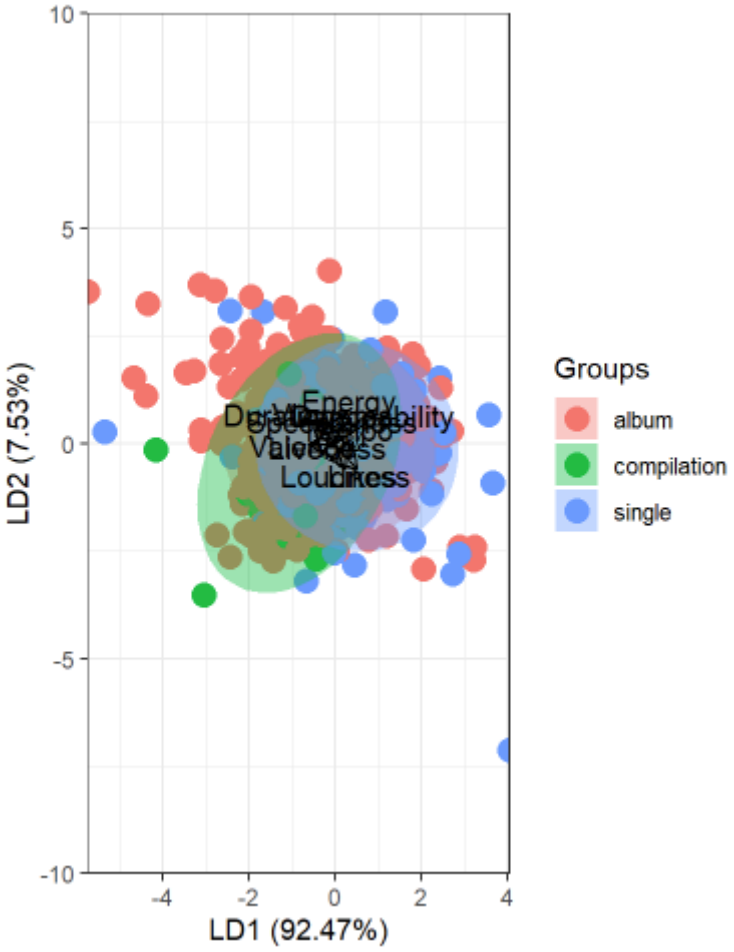

0.0
0.4
-6
-4
-2
0
2
4
group compilation


0.0
0.4
-6
-4
-2
0
2
4
group single


0.0
0.4
-6
-4
-2
0
2
4
group album


0.0
0.4
-6
-4
-2
0
2
4
group compilation


0.0
0.4
-6
-4
-2
0
2
4
group single




```
##              Actual
## Predicted     album compilation single
##   album        1056           44    353
##   compilation     0            1      0
##   single         21            0     24
```

total correct classification is 1056+1+24=1081

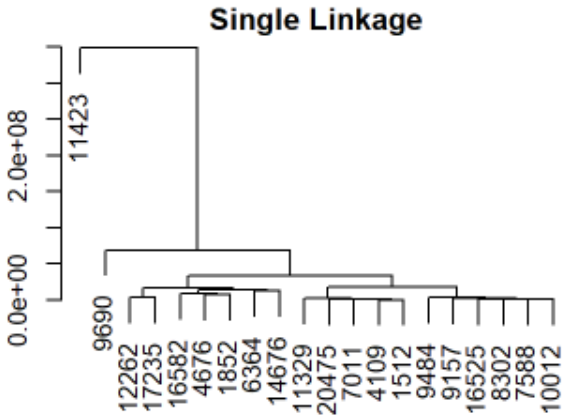
```
## [1] 0.7211474
```

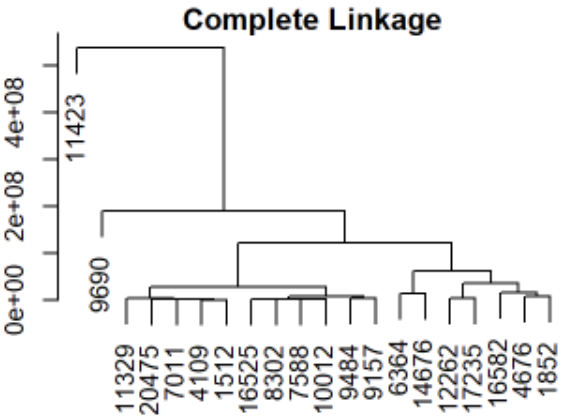
The accuracy of the model is 72.1%. It is not enough. The classification error rate is 1-0.72= 0.28

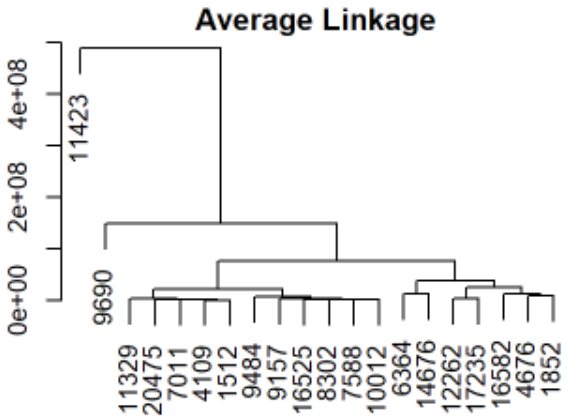
```
##              Actual
## Predicted     album compilation single
##   album         252           13     99
##   compilation     1            0      0
##   single          7            0      5
```

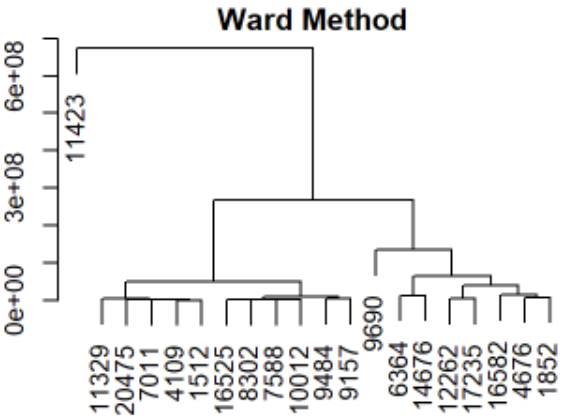
```
## [1] 0.6816976
```

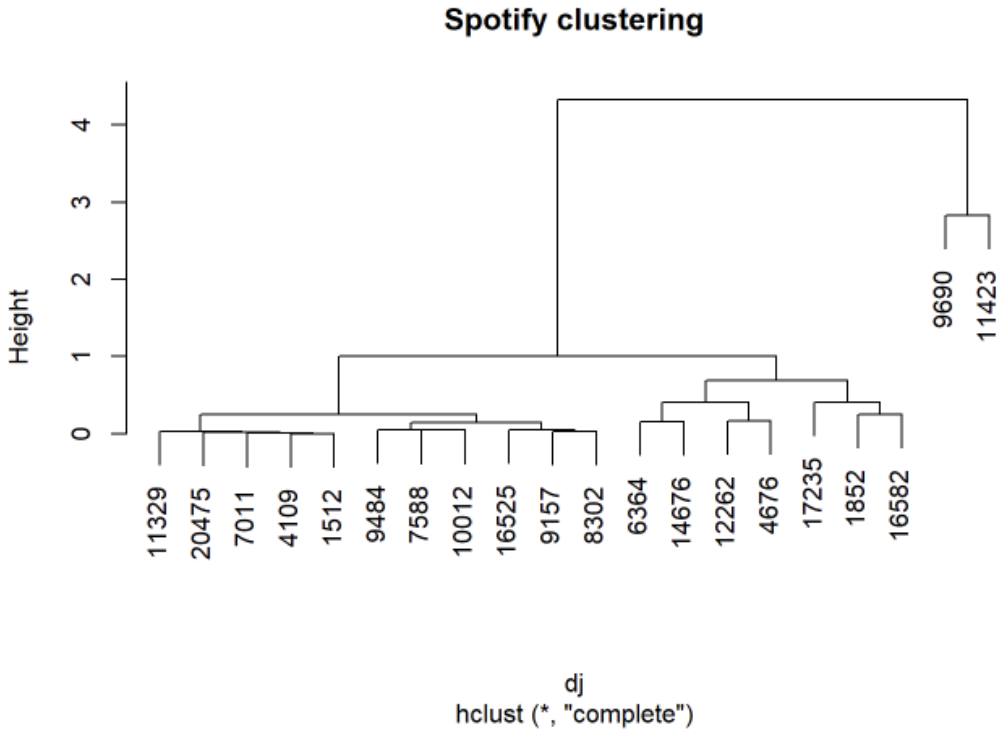
The accuracy of model is around 68.1%. So the models corretly classifies songs by album types with 68.1% for the test data.


Single Linkage
0.0e+00
2.0e+08
11423
9690
12262
17235
16582
4676
1852
6364
14676
11329
20475
7011
4109
1512
9484
9157
16525
8302
7588
10012

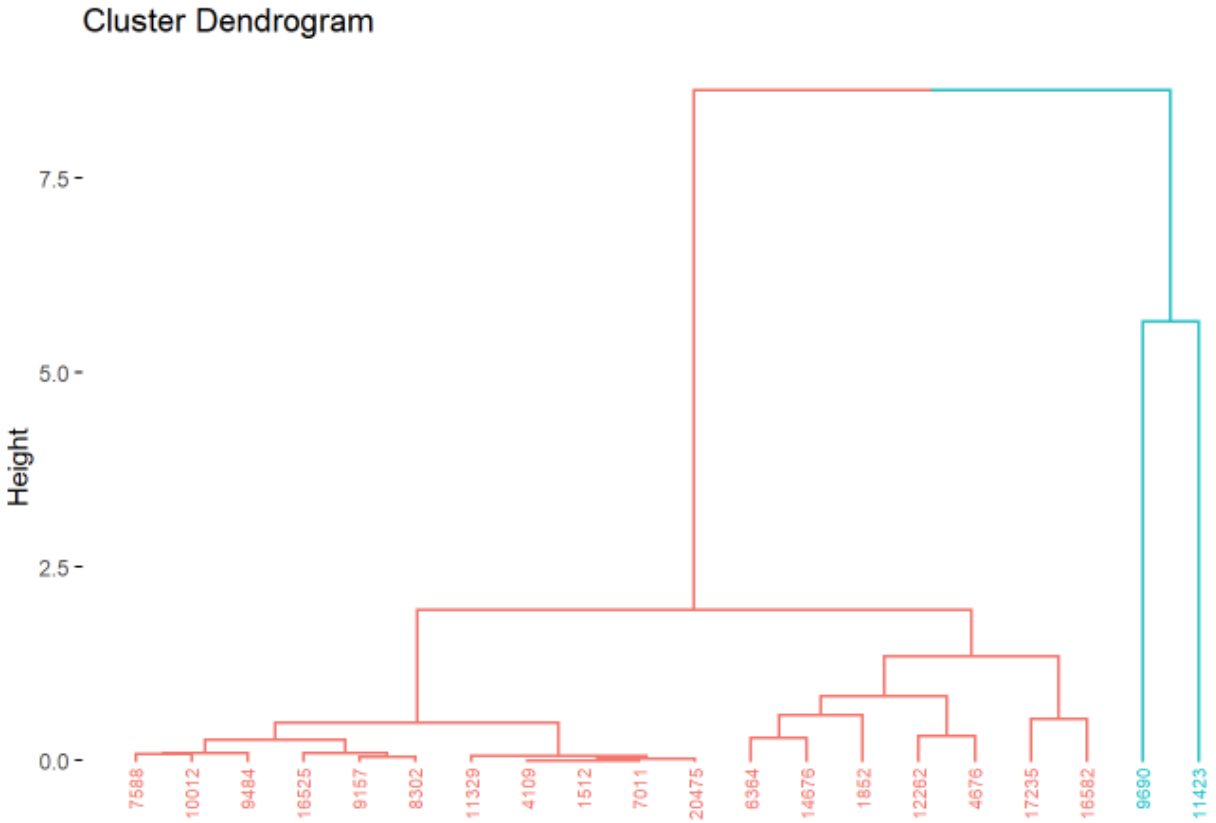

Complete Linkage
0e+00
2e+08
4e+08
11423
9690
11329
20475
7011
4109
1512
16525
8302
7588
10012
9484
9157
6364
14676
12262
17235
16582
4676
1852

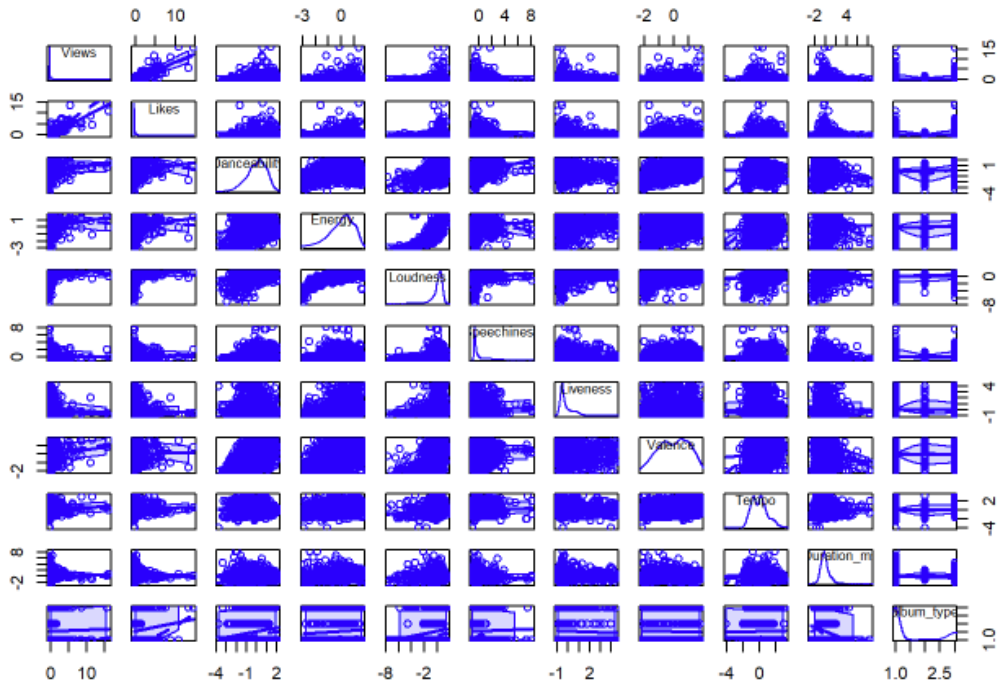

Average Linkage
0e+00
2e+08
4e+08
11423
9690
11329
20475
7011
4109
1512
9484
9157
16525
8302
7588
10012
6364
14676
12262
17235
16582
4676
1852


Ward Method
0e+00
3e+08
6e+08
11423
11329
20475
7011
4109
1512
16525
8302
7588
10012
9484
9157
9690
6364
14676
12262
17235
16582
4676
1852

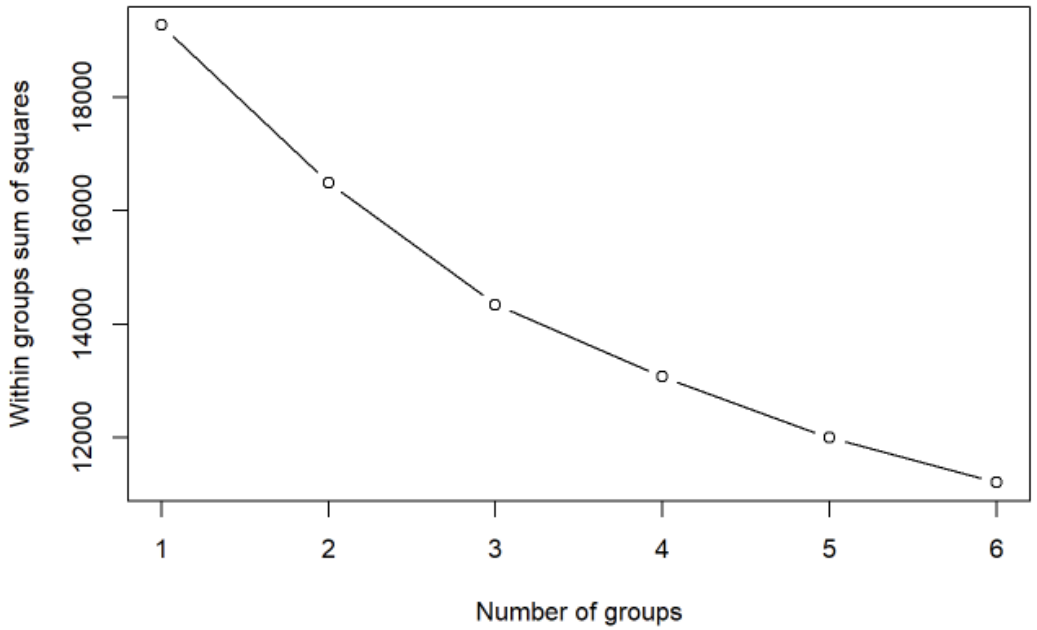



```
##
## Call:
## hclust(d = dj)
##
## Cluster method   : complete
## Distance         : euclidean
## Number of objects: 20
```




```
##        Views        Likes Danceability       Energy     Loudness  Speechiness
##    1.0000000    1.0000000    1.0000000    1.0000000    1.0000000    1.0000000
##     Liveness      Valence        Tempo  Duration_ms   album_type
##    1.0000000    1.0000000    1.0000000    1.0000000    0.2789421
```




3 is the elbow points

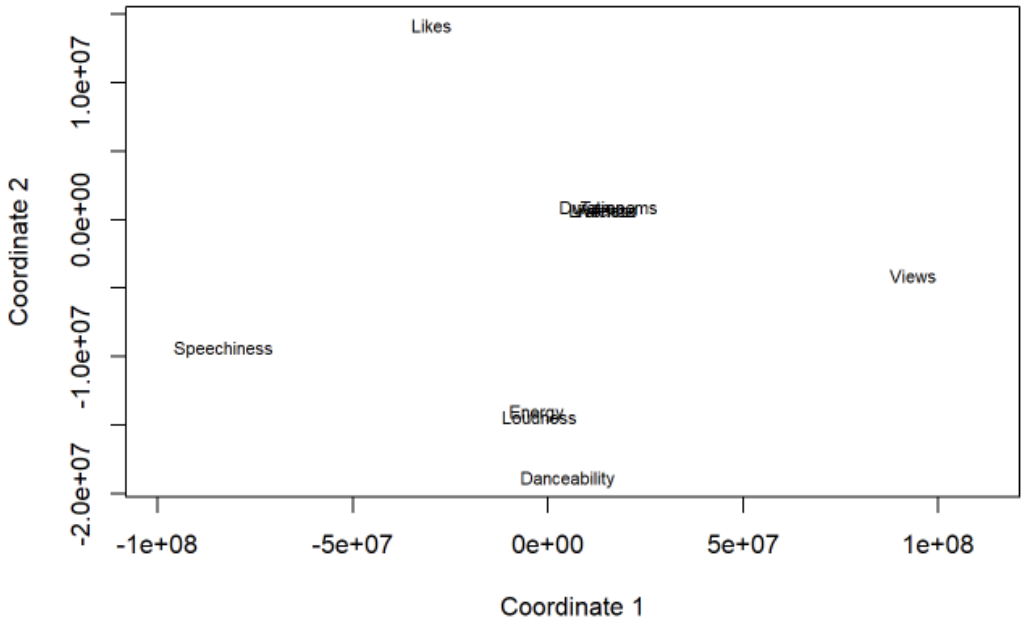
```
##       Views      Likes Danceability      Energy  Loudness Speechiness
## 1 -1.793480 -0.2576669    3.3208840   1.4038507  2.837351   17.614731
## 2  1.569420  0.9426030    0.6714779   0.8789135  2.098868   -2.848021
## 3 -1.210602 -2.1367993   -6.3092749 -20.8539556 -5.012630   -2.652840
##     Liveness     Valence      Tempo Duration_ms album_type
## 1  0.3738808   0.3887161  0.2273580  -2.2414527   9.543000
## 2  0.3222040   5.2344770  0.3674183  -0.1210294  14.702211
## 3 -2.0105576 -14.6050263 -2.1675313   0.9125343   2.575406
```

```
## Warning in cmdscale(cluster_sample2, k = 9, eig = TRUE): only 6 of the first 9
## eigenvalues are > 0
```

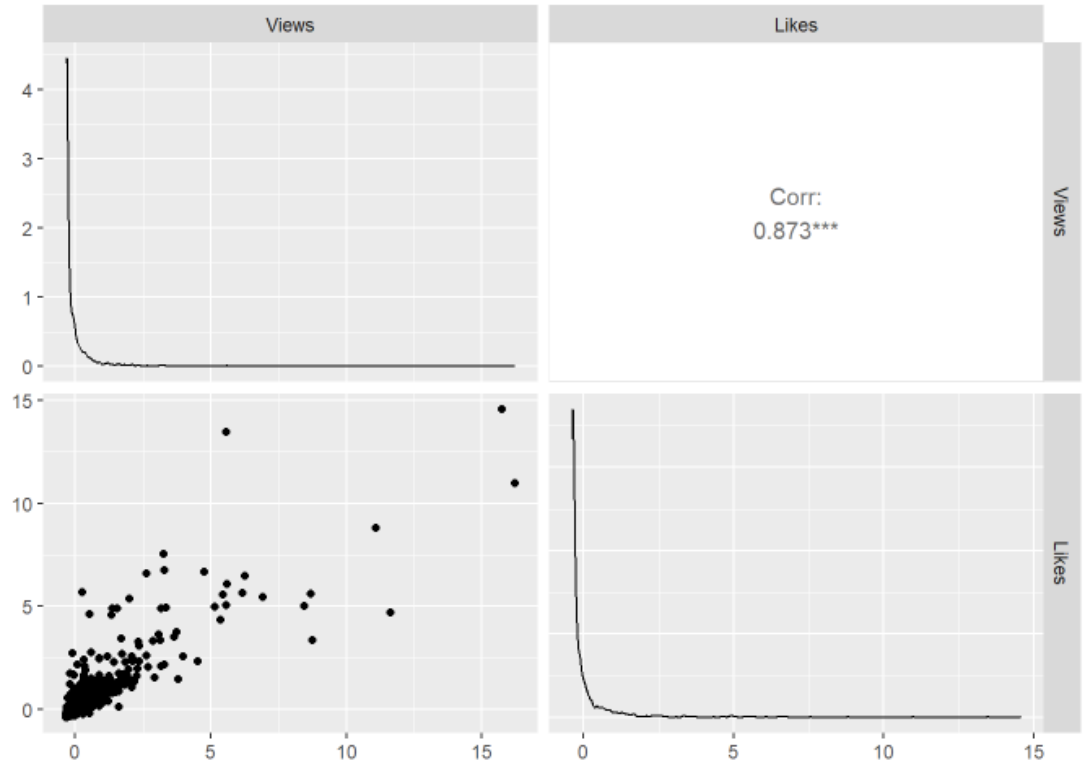
```
##  [1]  1.748489e+16  1.074170e+15  2.051104e+13  8.141297e+12  8.851645e+11
##  [6]  1.760026e+11 -1.125348e+14 -6.826224e+14 -1.241119e+15 -1.376317e+16
```

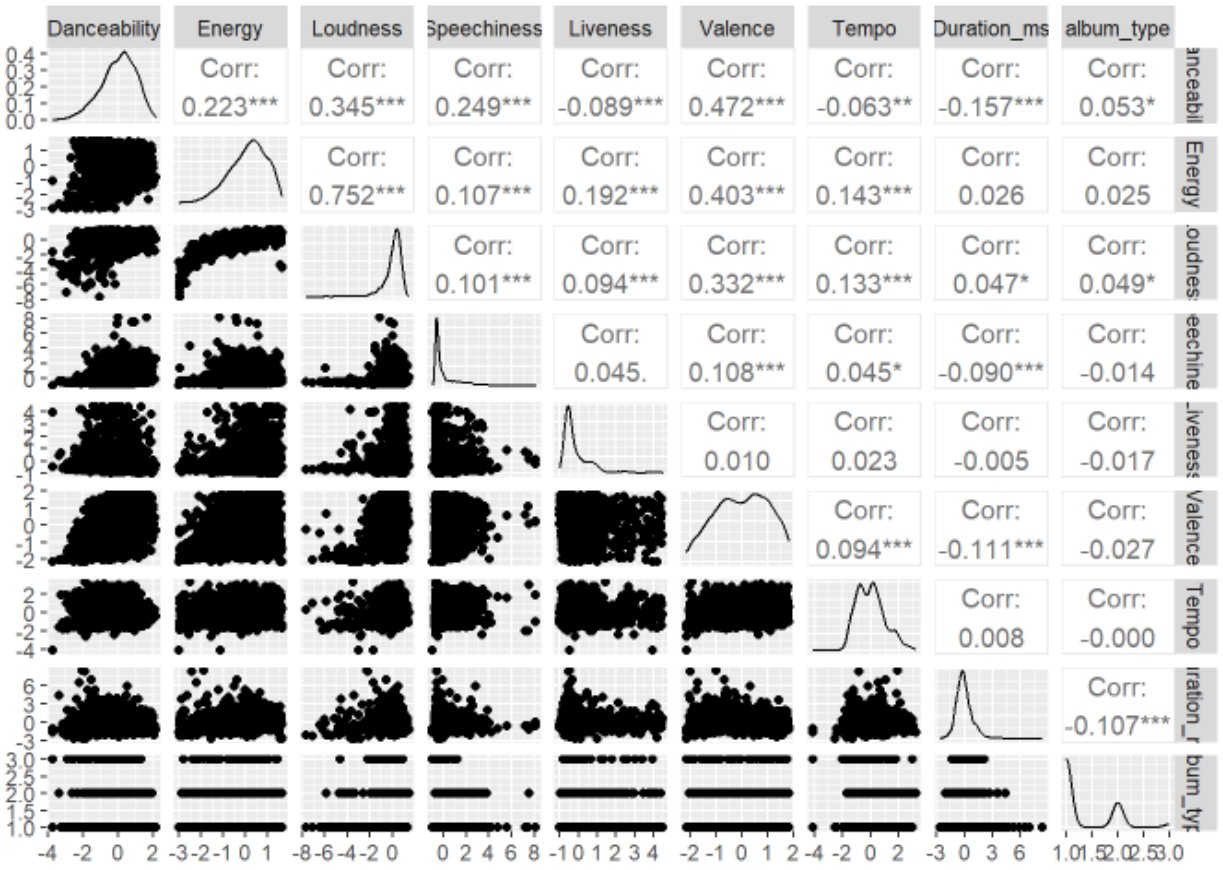
```
##  [1] 0.5084558 0.5396923 0.5402888 0.5405255 0.5405513 0.5405564 0.5438289
##  [8] 0.5636793 0.5997707 1.0000000
```

```
##  [1] 0.6135046 0.6158200 0.6158209 0.6158210 0.6158210 0.6158210 0.6158464
##  [8] 0.6167815 0.6198726 1.0000000
```



```
## album_type
##    1    2    3
## 1337  481   58
```



```
## $Xcor
##            Views     Likes
## Views  1.0000000 0.8728789
## Likes  0.8728789 1.0000000
##
## $Ycor
##              Danceability     Energy   Loudness Speechiness     Liveness
## Danceability   1.00000000 0.22287732 0.34468087  0.24947149 -0.089435804
## Energy         0.22287732 1.00000000 0.75212459  0.10741110  0.191544053
## Loudness       0.34468087 0.75212459 1.00000000  0.10062971  0.093689883
## Speechiness    0.24947149 0.10741110 0.10062971  1.00000000  0.044733254
## Liveness      -0.08943580 0.19154405 0.09368988  0.04473325  1.000000000
## Valence        0.47226099 0.40321255 0.33194829  0.10766200  0.010160235
## Tempo         -0.06281959 0.14291261 0.13323714  0.04528553  0.022823795
## Duration_ms   -0.15668362 0.02611259 0.04699727 -0.08972338 -0.005078871
## album_type     0.05286414 0.02459932 0.04862961 -0.01439931 -0.016945282
##                   Valence         Tempo  Duration_ms    album_type
## Danceability  0.47226099 -0.0628195933 -0.156683617  0.0528641374
## Energy        0.40321255  0.1429126139  0.026112591  0.0245993228
## Loudness      0.33194829  0.1332371373  0.046997274  0.0486296125
## Speechiness   0.10766200  0.0452855308 -0.089723382 -0.0143993089
## Liveness      0.01016024  0.0228237948 -0.005078871 -0.0169452815
## Valence       1.00000000  0.0935943498 -0.111093048 -0.0270189651
## Tempo         0.09359435  1.0000000000  0.007679246 -0.0003911746
## Duration_ms  -0.11109305  0.0076792464  1.000000000 -0.1072716712
## album_type   -0.02701897 -0.0003911746 -0.107271671  1.0000000000
```

```
## $XYcor
##                      Views       Likes Danceability     Energy   Loudness
## Views         1.000000000  0.87287889   0.09827919 0.08745398 0.12600815
## Likes         0.872878885  1.00000000   0.10234528 0.08852479 0.14523776
## Danceability  0.098279195  0.10234528   1.00000000 0.22287732 0.34468087
## Energy        0.087453975  0.08852479   0.22287732 1.00000000 0.75212459
## Loudness      0.126008150  0.14523776   0.34468087 0.75212459 1.00000000
## Speechiness   0.009066823  0.04078803   0.24947149 0.10741110 0.10062971
## Liveness     -0.026314067 -0.02762089  -0.08943580 0.19154405 0.09368988
## Valence       0.056422899  0.01567925   0.47226099 0.40321255 0.33194829
## Tempo         0.054570363  0.04992230  -0.06281959 0.14291261 0.13323714
## Duration_ms   0.031581646  0.01318804  -0.15668362 0.02611259 0.04699727
## album_type   -0.004381202  0.03106988   0.05286414 0.02459932 0.04862961
##                Speechiness     Liveness     Valence          Tempo  Duration_ms
## Views         0.009066823 -0.026314067  0.05642290  0.0545703634  0.031581646
## Likes         0.040788029 -0.027620886  0.01567925  0.0499222953  0.013188041
## Danceability  0.249471494 -0.089435804  0.47226099 -0.0628195933 -0.156683617
## Energy        0.107411097  0.191544053  0.40321255  0.1429126139  0.026112591
## Loudness      0.100629707  0.093689883  0.33194829  0.1332371373  0.046997274
## Speechiness   1.000000000  0.044733254  0.10766200  0.0452855308 -0.089723382
## Liveness      0.044733254  1.000000000  0.01016024  0.0228237948 -0.005078871
## Valence       0.107661997  0.010160235  1.00000000  0.0935943498 -0.111093048
## Tempo         0.045285531  0.022823795  0.09359435  1.0000000000  0.007679246
## Duration_ms  -0.089723382 -0.005078871 -0.11109305  0.0076792464  1.000000000
## album_type   -0.014399309 -0.016945282 -0.02701897 -0.0003911746 -0.107271671
##                 album_type
## Views        -0.0043812019
## Likes         0.0310698833
## Danceability  0.0528641374
## Energy        0.0245993228
## Loudness      0.0486296125
## Speechiness  -0.0143993089
## Liveness     -0.0169452815
## Valence      -0.0270189651
## Tempo        -0.0003911746
## Duration_ms  -0.1072716712
## album_type    1.0000000000
```

```
## [1] 0.1836423 0.1295465
```

```
## $xcoef
##            [,1]      [,2]
## Views  0.684202 -1.931863
## Likes -1.539853  1.352435
##
## $ycoef
##                        [,1]        [,2]
## Danceability -0.4314270201 -0.30079985
## Energy        0.0914661029 -0.11105923
## Loudness     -0.8096211697 -0.01434253
## Speechiness  -0.1889310072  0.41755705
## Liveness      0.1572095325  0.09686818
## Valence       0.5457077423 -0.53746855
## Tempo        -0.1937711883 -0.26164532
## Duration_ms  -0.0006031122 -0.35756607
## album_type   -0.3831680444  0.68863231
```

```
## $corr.X.xscores
##             [,1]       [,2]
## Views -0.6599028 -0.7513510
## Likes -0.9426271 -0.3338474
##
## $corr.Y.xscores
##                        [,1]        [,2]
## Danceability -0.090353829 -0.05144659
## Energy       -0.076478947 -0.04922507
## Loudness     -0.137429717 -0.04700585
## Speechiness  -0.056604014  0.03764729
## Liveness      0.024527956  0.01347972
## Valence       0.014460925 -0.08779613
## Tempo        -0.039535824 -0.03790580
## Duration_ms   0.001300586 -0.04317544
## album_type   -0.050840668  0.05048387
##
## $corr.X.yscores
##             [,1]        [,2]
## Views -0.1211861 -0.09733488
## Likes -0.1731062 -0.04324876
##
## $corr.Y.yscores
##                        [,1]       [,2]
## Danceability -0.492009811 -0.3971284
## Energy       -0.416455980 -0.3799799
## Loudness     -0.748355328 -0.3628493
## Speechiness  -0.308229666  0.2906083
## Liveness      0.133563738  0.1040531
## Valence       0.078745052 -0.6777191
## Tempo        -0.215287096 -0.2926039
## Duration_ms   0.007082171 -0.3332814
## album_type   -0.276846126  0.3896969
```

```
## Wilks' Lambda, using F-approximation (Rao's F):
##               stat   approx df1  df2      p.value
## 1 to 2:  0.9500592 5.376671  18 3730 1.320721e-12
## 2 to 2:  0.9832177 3.981285   8 1866 1.083789e-04
```

```
##  Hotelling-Lawley Trace, using F-approximation:
##                stat   approx df1  df2      p.value
## 1 to 2:  0.05197029 5.381812  18 3728 1.271983e-12
## 2 to 2:  0.01706875 3.981285   8 3732 1.036489e-04
```

```
##  Pillai-Bartlett Trace, using F-approximation:
##                stat   approx df1  df2      p.value
## 1 to 2:  0.05050680 5.371521  18 3732 1.371347e-12
## 2 to 2:  0.01678229 3.951826   8 3736 1.141128e-04
```

```
##  Roy's Largest Root, using F-approximation:
##                stat   approx df1  df2      p.value
## 1 to 1:  0.03372451 7.236253   9 1866 2.124554e-10
##
##  F statistic for Roy's Greatest Root is an upper bound.
```

```
##           [,1]      [,2]
## [1,]  0.684202 -1.931863
## [2,] -1.539853  1.352435
```

```
##                [,1]        [,2]
##  [1,] -0.4314270201 -0.30079985
##  [2,]  0.0914661029 -0.11105923
##  [3,] -0.8096211697 -0.01434253
##  [4,] -0.1889310072  0.41755705
##  [5,]  0.1572095325  0.09686818
##  [6,]  0.5457077423 -0.53746855
##  [7,] -0.1937711883 -0.26164532
##  [8,] -0.0006031122 -0.35756607
##  [9,] -0.2023701025  0.36370098
```

my data was already standardized so it didn't change.

```
## Wilks' Lambda, using F-approximation (Rao's F):
##                stat   approx df1  df2      p.value
## 1 to 2:  0.9500592 5.376671  18 3730 1.320721e-12
## 2 to 2:  0.9832177 3.981285   8 1866 1.083789e-04
```

```
##           [,1]      [,2]
## Views  0.684202 -1.931863
## Likes -1.539853  1.352435
```

```
##                      [,1]        [,2]
## Danceability -0.4314270201 -0.30079985
## Energy        0.0914661029 -0.11105923
## Loudness     -0.8096211697 -0.01434253
## Speechiness  -0.1889310072  0.41755705
## Liveness      0.1572095325  0.09686818
## Valence       0.5457077423 -0.53746855
## Tempo        -0.1937711883 -0.26164532
## Duration_ms  -0.0006031122 -0.35756607
## album_type   -0.3831680444  0.68863231
```