SLTC
Research University

Statistical Analysis of Spotify Songs Audio Features Dataset

EMA3200: Applied Statistics
Group Assignment

Group Members:

22UG1-0041-D.B.Yatigammana
22UG1-0819-H.M.N.V.Herath
22UG1-0481-S.M.C.Sankalpana
22UG1-0824-K.M.L.Madushan
22UG1-0477- N.K.B.N.Silva
22UG1-0883-D.K.Nivedya

Contents

1. Introduction

In the evolving landscape of music analytics, data-driven techniques have become essential for uncovering trends and patterns in popular music. This project applies advanced statistical methods to explore audio characteristics of spotify charting songs, aiming to extract meaningful insights about what makes a song resonate with audiences.

Leveraging a comprehensive dataset containing audio features such as **danceability, energy, loudness, key, tempo, and popularity**, this analysis utilizes both classical and modern statistical tools to visualize, model, and cluster songs based on their intrinsic attributes. By examining categorical and quantitative variables through visual exploration, regression modeling, and multivariate techniques, we aim to develop a deeper understanding of the audio landscape of successful songs.

The objective is to enhance interpretability of audio feature interactions and identify patterns that potentially contribute to a song's popularity. Through this, we not only demonstrate technical proficiency in applied statistics but also draw actionable insights for fields such as music recommendation, audio production, and entertainment analytics.

Objective of the Analysis:

- To explore and visualize relationships between song attributes.
- To build regression models to predict features such as popularity and energy.
- To apply multivariate analysis methods such as PCA, factor analysis, and clustering to reduce dimensions and group similar songs.

2. Proper Citation of Sources and Dataset Used

This report analyzes a dataset comprising **29479 records and 22 attributes**, capturing the audio features of songs that have appeared on the Billboard Hot 100 charts between the years 2000 and 2019. The dataset provides a comprehensive view of the musical characteristics that may contribute to a song’s popularity and chart performance.

The data also incorporates Spotify’s detailed audio analysis, collected through the Spotify Web API, providing a rich blend of musical metadata.

Key Features in the Dataset Include:

- **Identifiers and Metadata**:

- *song_id* (character) – Unique ID for the song.
- *performer* (character) – Name of the performing artist.
- *song* (character) – Title of the song.
- *spotify_genre* (character) – Genre classification from Spotify.
- *spotify_track_id* (character) – Spotify track ID.
- *spotify_track_preview_url* (character) – Link to the track preview.
- *spotify_track_album* (character) – Album name.

- **Audio Features** (Numerical - double):
  - *danceability* – Suitability of a track for dancing based on rhythm and beat stability.
  - *energy* – Perceptual measure of track intensity and activity.
  - *key* – Estimated musical key using standard Pitch Class notation (0 = C, 1 = C♯/D♭, etc.).
  - *loudness* – Average loudness of the track in decibels (dB).
  - *mode* – Major (1) or minor (0) modality of the track.
  - *speechiness* – Presence of spoken words in a track.
  - *acousticness* – Confidence measure of whether the track is acoustic.
  - *instrumentalness* – Likelihood that a track contains no vocal content.
  - *liveness* – Probability that the track was recorded in a live setting.
  - *valence* – Measure of musical positiveness (happiness vs sadness).
  - *tempo* – Estimated speed of the track in beats per minute (BPM).
  - *time_signature* – Estimated number of beats per bar (musical time signature).
  - *spotify_track_popularity* – Popularity score of the track on Spotify.

- **Additional Variables**:

- *spotify_track_duration_ms* (double) – Track duration in milliseconds.
- *spotify_track_explicit* (logical) – Whether the track is explicit (1) or not (0).

Variable Types in the Dataset:

- **Integer (int)**: Representing attributes like *year* and *duration (ms)*.
- **Categorical (object/character)**: Representing musical features such as *key*, *mode*, *genre*, and *album name*.
- **Float (double)**: Representing numerical values like *energy*, *tempo*, *danceability*, etc.
- **Logical (boolean)**: Representing explicit content indicator.

Dataset Completeness:

An important feature of this dataset is its robustness, with no missing values detected in key audio attributes. This completeness ensures high reliability for exploratory data analysis, predictive modeling, and trend forecasting tasks.

By examining these attributes, we can uncover insights into what defines successful Billboard-charting songs, track genre-based trends over time, and analyze the evolving landscape of popular music using quantitative audio features.

Citation:

- *Billboard Songs Audio Features Dataset*, retrieved from Data.World (originally sourced from Billboard.com and Spotify).
- *Billboard Hot 100 - Wikipedia*: Wikipedia Billboard Top 100

3.Data Preparation

Data Cleaning Steps:

- Checked and handled missing values (if any).

- Converted datatypes where necessary (e.g., categorical variables as factors).
- Standardized feature names.

Overview of Key Variables:

- *Categorical Variables:* key, mode
- *Quantitative Variables:* danceability, energy, loudness, valence, tempo, popularity

R Code + Output:

```
# Load the dataset (modify the path accordingly)
data<- read.csv("C:\\Users\\DELL\\Desktop\\audio_features (3).csv")
head(data,5)

#understand the data set
colnames(data)
ncol(data)
nrow(data)
summary(data)
str(data)

#clean the data set
sum(duplicated(data)) # duplication rows
data <- data[!duplicated(data), ]
sum(duplicated(t(data))) # duplication column
sum(is.na(data)) # check missing value in dataset
filtered_data<- na.omit(data) #remove rows with missing values

# Convert character to factor (only for categorical variables)
filtered_data <- filtered_data %>%
  mutate(across(where(is.character), as.factor))

# Convert TRUE/FALSE to 1/0
filtered_data$spotify_track_explicit <- as.numeric(filtered_data$spotify_track_explicit)
```

4. Data Visualization

In this section, we explore the Billboard Top 100 Songs dataset through various types of visualizations, categorized into univariate, bivariate, and multivariate analyses.

Univariate Analysis

Univariate analysis involves examining one variable at a time to understand its distribution and characteristics.

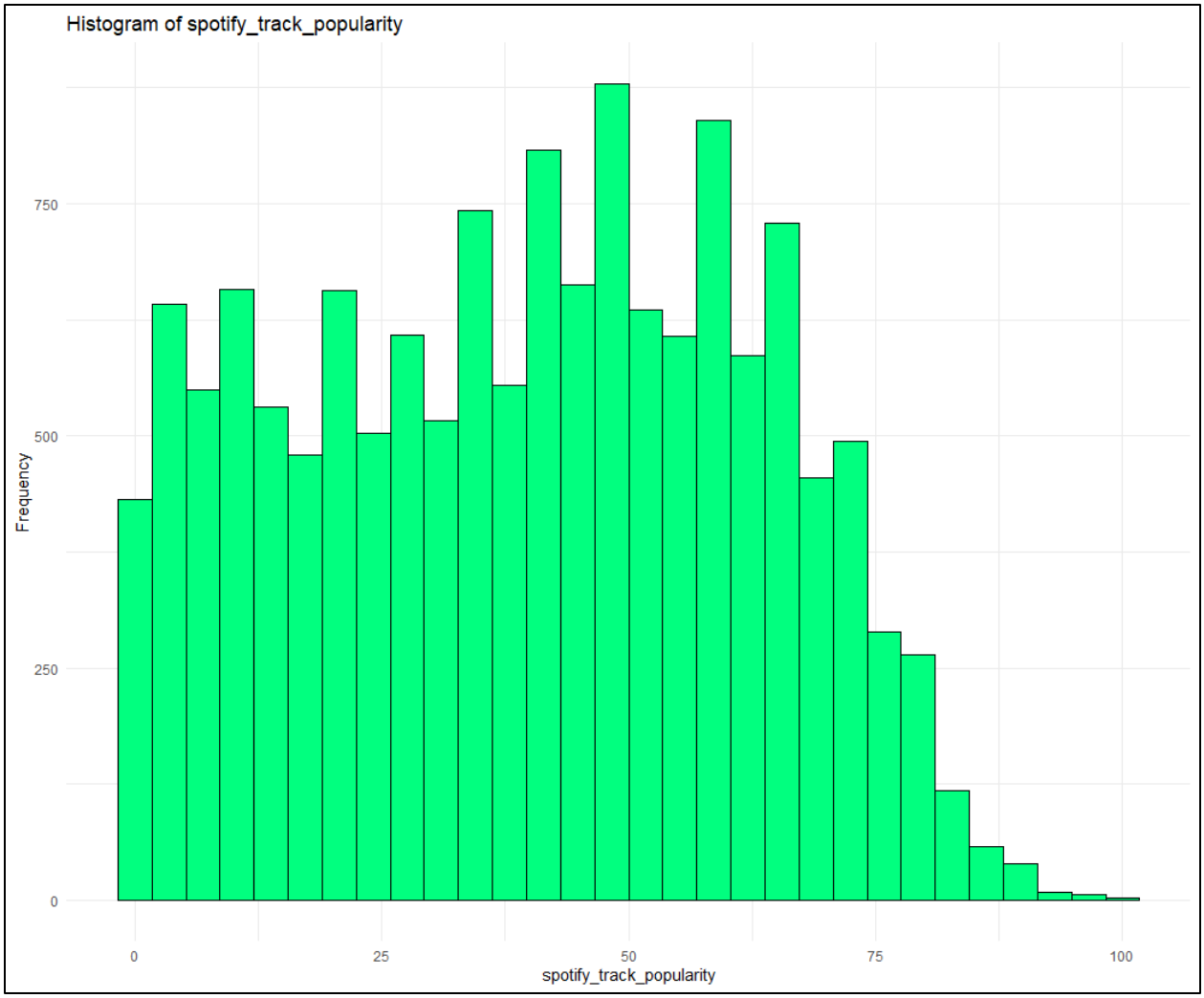
- **Histogram**:
    - We plotted histograms for key numerical features such as **Energy**, **Danceability**, **Tempo**, **Loudness**, and **Valence**.

    - Histograms help to visualize the distribution of values and identify patterns like skewness, normality, or the presence of outliers.

Histogram of Energy:
To explore the distribution of the spotify_track_popularity feature, a histogram was created using ggplot2:

R Code + Output:

```
# Universite Visualization
ggplot(filtered_data, aes(x = spotify_track_popularity)) +
  geom_histogram(bins = 30, fill = "springgreen", color = "black") +
  labs(title = "Histogram of spotify_track_popularity", x = "spotify_track_popularity", y = "Frequency") +
  theme_minimal()
```



Interpretation of the Histogram (spotify_track_popularity):
X-axis (spotify_track_popularity):
Represents the popularity score of Spotify tracks (from 0 to 100).

Y-axis (Frequency):
Shows how many tracks fall into each popularity range (how many times each popularity score appears).

Observations:
Most tracks have a popularity between 20 and 70.
You can see a thick, consistent block of bars between 20 and 70 on the x-axis.

Very few tracks are extremely popular (above 80).
After around 75, the height of bars drops quickly — meaning only a small number of tracks are super popular.

Popularity is not normally distributed.

It's right-skewed (a long tail towards the right).
Most tracks have medium or low popularity.
Very few tracks achieve extremely high popularity (close to 100).

Peak/Mode around 40–60:
There are more tracks that have popularity scores around 40, 50, and 60 compared to other areas.

"Most Spotify tracks have moderate popularity (20-70), very few are highly popular (above 80), and the distribution is right-skewed."

Bivariate Analysis

Bivariate analysis explores the relationship between two variables. We considered different combinations of categorical and quantitative variables:

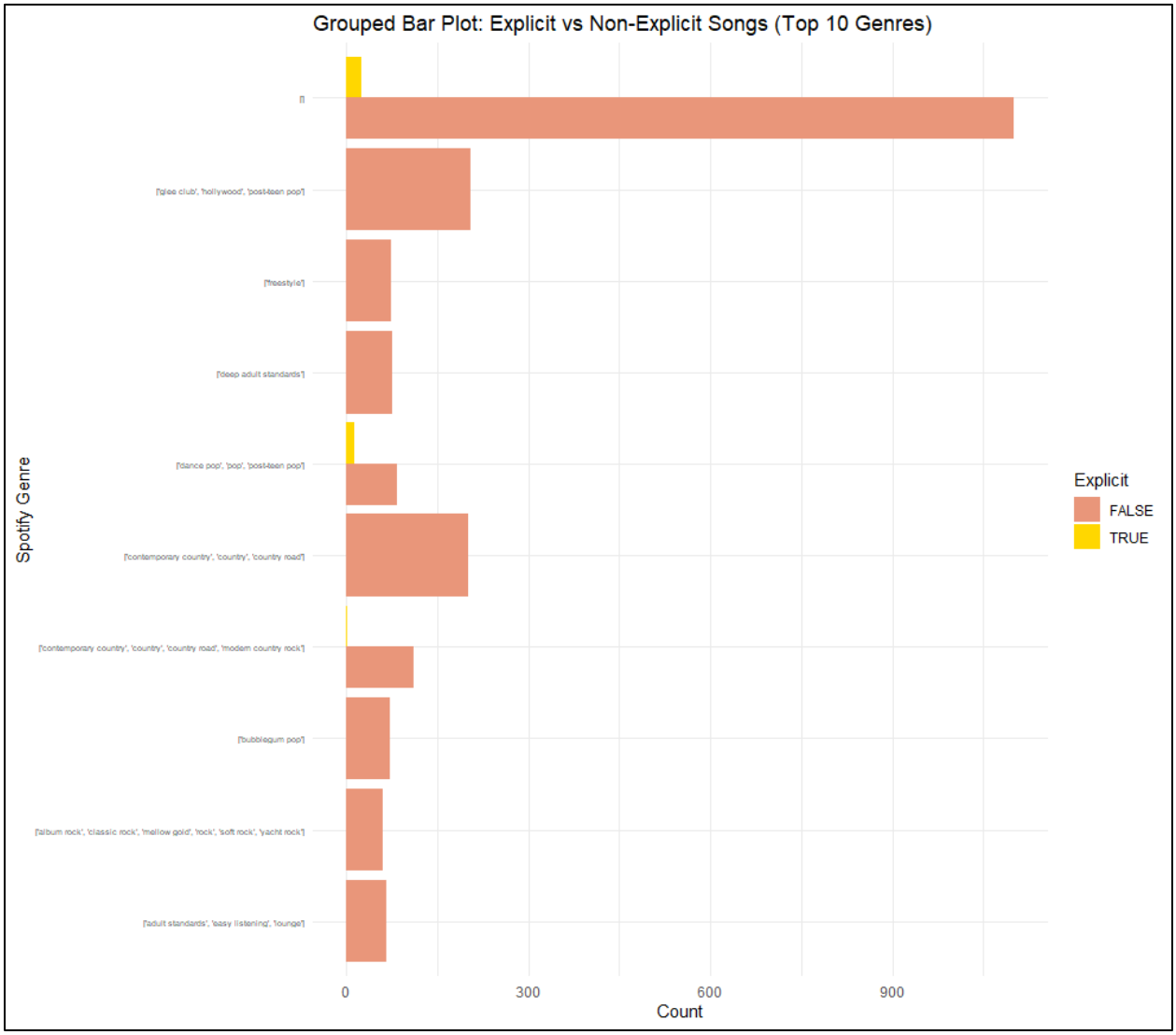
Categorical vs. Categorical

- **Explicit vs Non-Explicit Songs by Genre (Grouped Bar Plot)**:
    - A grouped bar plot was created to compare the number of explicit and non-explicit songs across different genres.
    - This visualization highlights which genres are more likely to contain explicit content.

R Code + Output:

```
# Get top 10 genres by count
top10_genres <- filtered_data %>%
  count(spotify_genre, sort = TRUE) %>%
  slice_head(n = 10) %>%
  pull(spotify_genre)

# Filter data for top 10 genres
filtered_top10 <- filtered_data %>%
  filter(spotify_genre %in% top10_genres)

# Plot
ggplot(filtered_top10, aes(x = spotify_genre, fill = factor(spotify_track_explicit), group =
spotify_track_explicit)) +
  geom_bar(position = "dodge") +
  labs(
          title = "Grouped Bar Plot: Explicit vs Non-Explicit Songs (Top 10 Genres)",
          x = "Spotify Genre",
```

```
        y = "Count",
        fill = "Explicit"
) +
coord_flip() +
scale_fill_manual(values = c("darksalmon", "gold")) +
theme_minimal() +
theme (axis.text.y = element_text(size = 5))
```



The output of this code will be a **grouped bar plot** showing the distribution of explicit and non-explicit songs across different Spotify genres.

1. **X-axis (Spotify Genre):** Each bar group represents a different Spotify genre, such as "Pop," "Rock," "Hip-Hop," etc.

2. **Y-axis (Count):** The height of the bars represents the count (or frequency) of songs within each genre. This will show how many songs fall under the explicit and non-explicit categories for each genre.

3. **Bars (Explicit vs Non-Explicit):**

    - **Explicit songs** will be represented by one color (e.g., dark salmon), and

    - **Non-explicit songs** will be represented by another color (e.g., gold).

4. **Group Positioning (Dodge):** The bars for each genre will be placed side-by-side, showing a comparison between explicit and non-explicit songs within that genre.

5. **Flipped Coordinates (Horizontal bars):** The bars will be horizontal, making it easier to compare the values for different genres (especially if there are long genre names).

Interpretation:

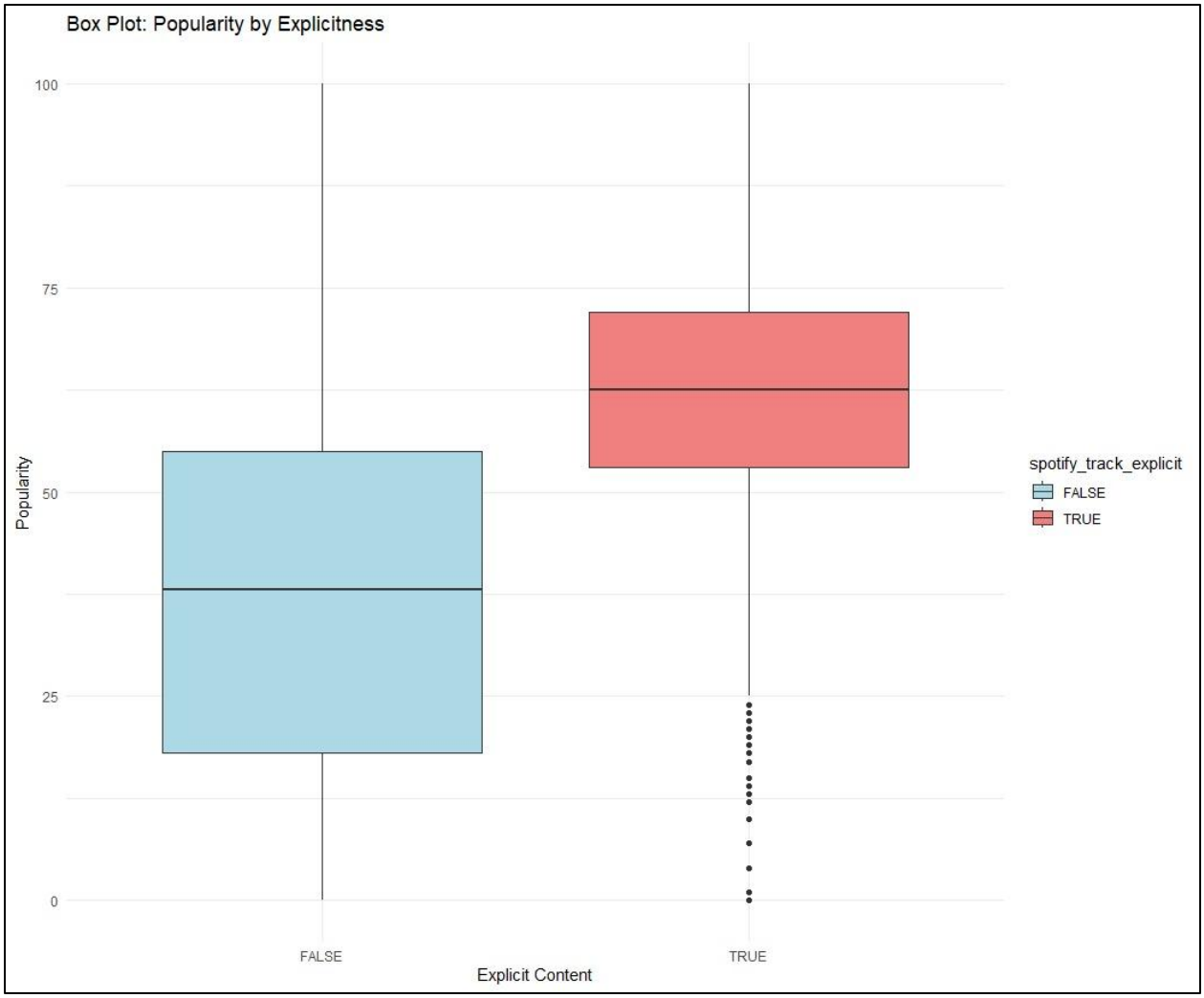
Based on the box plot, non-explicit songs generally have a higher median valence, suggesting they tend to be more cheerful or positive. Explicit songs show a wider range of valence, indicating more emotional variety , some are happy, others quite dark.

Categorical vs. Quantitative

- **Valence by Explicitness (Box Plot)**:
    - We used box plots to show the distribution of the **Valence** score (musical positiveness) for explicit and non-explicit tracks.
    - This helps to understand whether explicit songs tend to have more positive or negative emotions compared to non-explicit songs.

R Code + Output:

```
#Categorical vs quantitative -> Valence by Explicitness
#Box plot
# Convert spotify_track_explicit to a factor to use as x-axis
filtered_data$spotify_track_explicit <- factor(filtered_data$spotify_track_explicit)

ggplot(filtered_data, aes(x = spotify_track_explicit, y = spotify_track_popularity, fill =
spotify_track_explicit)) +
  geom_boxplot() +
  labs(
    title = "Box Plot: Popularity by Explicitness",
    x = "Explicit Content",
    y = "Popularity"
  ) +
  scale_fill_manual(values = c("lightblue", "lightcoral")) +
  theme_minimal()
```



Axes and Variables

- **X-axis:** Explicit Content (whether the song has explicit lyrics)
    - FALSE = Not explicit
    - TRUE = Explicit
- **Y-axis:** Popularity score (on a scale from 0 to 100)
- **Color:**
    - Light Blue = Not explicit (FALSE)
    - Light Coral = Explicit (TRUE)

Interpretation:

Explicit songs ("TRUE") are generally more popular!

- The median popularity (middle line inside the box) for explicit songs is noticeably higher than for non-explicit songs.
- Explicit songs' median is around 65, whereas non-explicit songs' median is closer to 40.

Spread of Popularity:

- The interquartile range (IQR = box height) for explicit songs is tighter and higher overall compared to non-explicit songs.
- Non-explicit songs have a wider range (the blue box is taller), meaning their popularity is more spread out and varied.

Outliers:

- Explicit songs (TRUE) show several outliers at the bottom (small dots).
  Some explicit songs are very unpopular (low scores) even though on average they do better.

Overall Popularity:

- Explicit tracks tend to cluster at higher popularity levels, while non-explicit tracks are more scattered across the full range (including very low popularity).

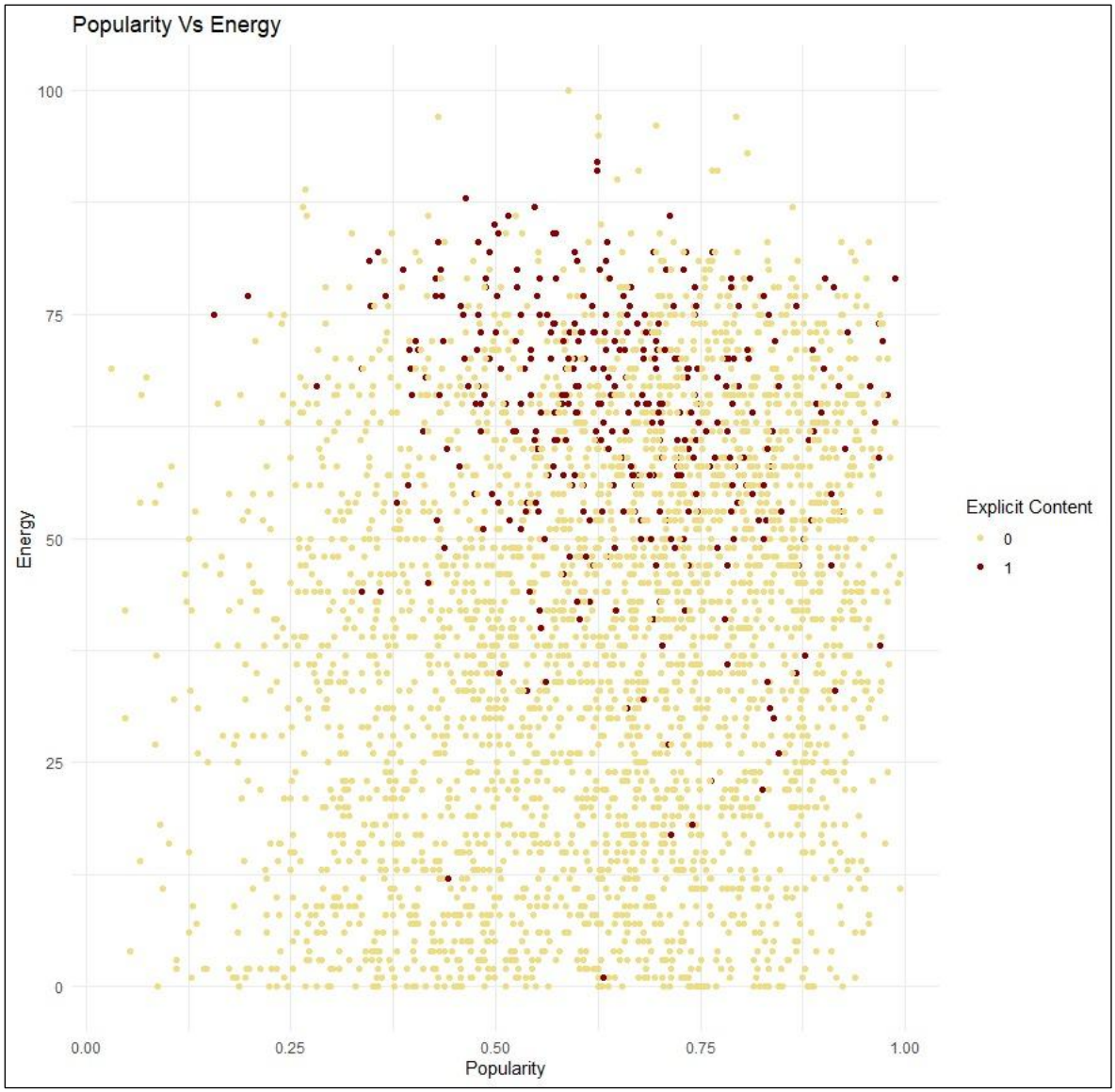
Quantitative vs. Quantitative

Popularity vs. Energy (Scatter Plot):

- A scatter plot was used to explore the relationship between Energy and Popularity.
- This visualization helps to observe how the energetic feel of a song influences its popularity on Spotify.
- Patterns and clusters in the plot can reveal whether high-energy tracks tend to be more popular or if there's no strong correlation between the two.

R Code + Output:

```
#quantitative vs quantitative -> popularity Vs energy
#scatter plot
ggplot(filtered_data[1:3500,], aes(x = energy, y = spotify_track_popularity, color =
spotify_track_explicit)) +
  geom_point() +
  labs(
    title = "Popularity Vs Energy",
    x = "Popularity",
    y = "Energy",
    color = "Explicit Content"
  ) +
  scale_color_manual(values = c("lightgoldenrod", "red4")) +
  theme_minimal()
```



Axes and Variables

- **X-axis:** Popularity (between 0 and 1)
- **Y-axis:** Energy (from 0 to 100)
- **Color** indicates **Explicit Content**:
    - Light Yellow (lightgoldenrod) = Non-explicit songs (FALSE)
    - Dark Red (red4) = Explicit songs (TRUE)

Interpretation:

Data Spread:

- Songs are spread across all energy levels (from low to high), but the majority cluster between 40 to 80 energy.

Popularity Range:

- Most songs have popularity between 0.3 and 0.8.
- Very few songs are close to popularity 1.0 (max).

Explicit vs Non-explicit:

- Explicit songs (dark red dots) tend to concentrate more where popularity is higher (around 0.5 - 0.8) and energy is moderately high (50-80).
- Non-explicit songs (light yellow dots) are spread wider across all energy levels and lower popularity too.

High Energy and Popularity:

- Songs with both high energy (>75) and high popularity (>0.7) are mostly explicit

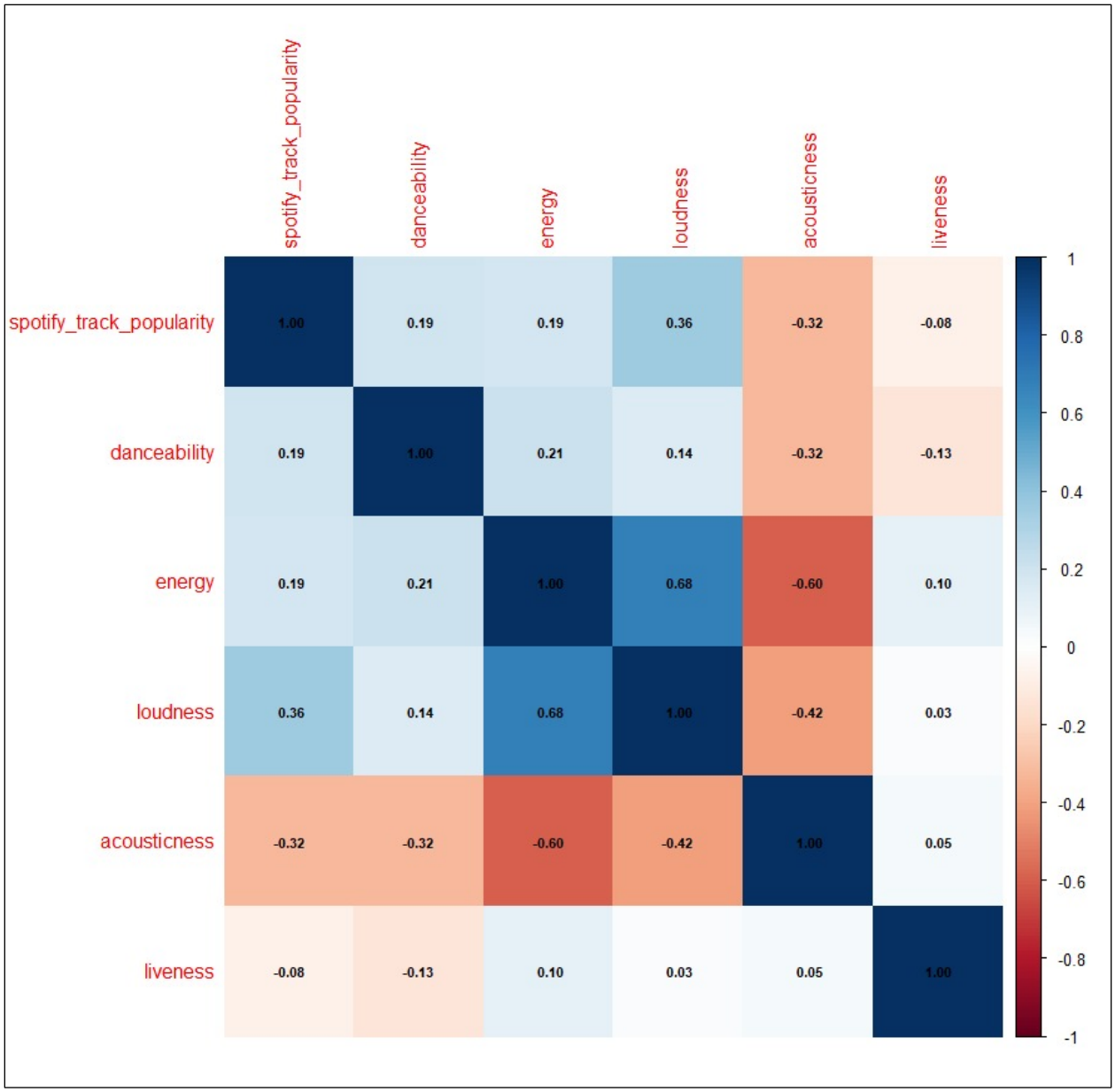
Multivariate Analysis

Multivariate analysis examines relationships involving more than two variables simultaneously.

- **Co-plot: Energy vs Tempo by Explicitness**:
    - A co-plot (conditional plot) was created with **Energy** as the dependent variable and **Tempo** as the independent variable, conditioned on the **Explicitness** of the song (explicit vs non-explicit).
    - This visualization allows for a deeper analysis of how tempo affects the energy of a song differently depending on whether the track is explicit or not.

R Code + Output:

```
#Multivarite  visuvaluzation

#install.packages("corrplot")
library(corrplot)
# Select numeric columns properly
num_vars <- filtered_data %>%
  select(spotify_track_popularity, danceability, energy, loudness, acousticness, liveness) %>%
  mutate(across(everything(), as.numeric))

# Now plot
corrplot(cor(num_vars), method = "color")
corrplot(cor(num_vars), method = "color", addCoef.col = "black", number.cex = 0.7)
```



Interpretation

A correlation analysis was conducted to explore the relationships among key musical attributes: spotify_track_popularity, danceability, energy, loudness, acousticness, and liveness. The Pearson correlation coefficients were computed and visualized using a color-coded correlation matrix.

The following key insights were identified:

Spotify Track Popularity shows a moderate positive correlation with loudness (r = 0.36), suggesting that louder tracks tend to be more popular. There are weak positive correlations with danceability (r = 0.19) and energy (r = 0.19), indicating that tracks with higher danceability and energy may experience slightly higher

popularity. A weak negative correlation with acousticness (r = -0.32) implies that more acoustic songs tend to have slightly lower popularity. The correlation with liveness is negligible (r = -0.08).

Energy is strongly positively correlated with loudness (r = 0.68), implying that more energetic tracks are generally louder. Additionally, energy exhibits a strong negative correlation with acousticness (r = -0.60), indicating that highly energetic tracks are typically less acoustic.

Loudness demonstrates a moderate negative correlation with acousticness (r = -0.42), further supporting the relationship that louder tracks tend to be less acoustic.

Danceability shows very weak correlations with all other variables, with the highest being a slight positive correlation with energy (r = 0.21).

Liveness does not show strong correlations with any other variable, suggesting it may be relatively independent of the other measured attributes in this dataset.

Overall, the analysis indicates that loudness and energy are closely related characteristics that positively influence track popularity, whereas acousticness tends to be negatively associated with both energy and popularity.

5. Regression Analysis (LO 02)

5.1 Simple Linear Regression

Predicting spotify_track_popularity from Loudness

```
# Simple Linear Model
 lm_simple <- lm(spotify_track_popularity ~ loudness, data = filtered_data)
 summary(lm_simple)
```

```
> # Simple Linear Model (e.g., Predicting Energy)
>  lm_simple <- lm(spotify_track_popularity ~ loudness, data = filtered_data)
>  summary(lm_simple)

Call:
lm(formula = spotify_track_popularity ~ loudness, data = filtered_data)

Residuals:
    Min      1Q  Median      3Q     Max
-60.817 -16.633   1.077  16.090  69.248

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 59.59806    0.46805  127.33   <2e-16 ***
loudness     2.27807    0.04953   45.99   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 21.19 on 14355 degrees of freedom
Multiple R-squared:  0.1284,    Adjusted R-squared:  0.1284
F-statistic:  2116 on 1 and 14355 DF,  p-value: < 2.2e-16

>
```

Simple Linear Model: Predicting Popularity from Loudness

Variable Selection Justification

The predictor variable **loudness** was selected to model spotify_track_popularity based on insights obtained from the coplot of the multivariate visualization.
Among the quantitative variables analyzed, loudness exhibited a relatively higher correlation with popularity, suggesting a meaningful linear relationship worth further investigation through simple linear regression.

Model Overview

- **Dependent Variable (Target)**: spotify_track_popularity
- **Independent Variable (Predictor)**: loudness

The model specification is:

Popularity = 59.598 + 2.278 × Loudness

This simple linear regression model seeks to quantify how variations in loudness impact a track's popularity.

Interpretation of Coefficients

Term	Estimate	Interpretation
Intercept	59.598	When the loudness is 0 dB, the predicted popularity score is approximately 59.6.
Loudness	2.278	For each 1 dB increase in loudness, the popularity score is expected to increase by approximately 2.28 points, assuming other factors remain constant.

Statistical Significance

The coefficient for loudness is highly statistically significant, with a p-value less than 2e-16.
This strong significance is further confirmed by the three asterisks (***) notation, indicating that the probability of this relationship being due to random chance is extremely low.

Model Fit Summary

Metric	Value	Interpretation
Multiple R-squared	0.1284	Approximately 12.84% of the variance in popularity is explained by loudness.
Adjusted R-squared	0.1284	The adjusted value remains consistent with R-squared, as only one predictor is involved.

Residual Standard Error	21.19	The average error in predicted popularity scores is approximately ±21 points.
F-statistic	2116 (p-value < 2.2e-16)	Indicates that the overall model is highly statistically significant.

Model Evaluation

This simple linear model is appropriate as both the dependent and independent variables are continuous and a statistically significant relationship exists between loudness and popularity.
Nevertheless, the relatively low R-squared value suggests that while loudness contributes to popularity, it explains only a small portion of the variability. Other factors such as danceability, energy, valence, genre, artist fame, and marketing are likely to play important roles in determining overall popularity.

Conclusion

Louder songs tend to be slightly more popular, and the relationship between loudness and popularity is statistically significant. However, loudness alone accounts for only a limited portion of the factors influencing a song's popularity.

5.2 Multiple Linear Regression Model

Predicting spotify_track_ Popularity with Loudness, Danceability, Valence ,energy

```
# Multiple Linear Model
lm_multiple <- lm(spotify_track_popularity ~ loudness + danceability + valence + energy, data = filtered_data)
 summary(lm_multiple)
```

```
> # Multiple Linear Model
>  lm_multiple <- lm(spotify_track_popularity ~ loudness + danceability + valence + energy, data = filtered_data)
>  summary(lm_multiple)

Call:
lm(formula = spotify_track_popularity ~ loudness + danceability +
    valence + energy, data = filtered_data)

Residuals:
    Min      1Q  Median      3Q     Max
-66.994 -15.117   0.404  14.957  71.047

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   47.41015    1.31720  35.993  < 2e-16 ***
loudness       1.86251    0.06726  27.689  < 2e-16 ***
danceability  40.94117    1.19000  34.404  < 2e-16 ***
valence      -32.40781    0.84959 -38.145  < 2e-16 ***
energy         5.97886    1.27595   4.686 2.81e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 19.84 on 14352 degrees of freedom
Multiple R-squared:  0.2362,    Adjusted R-squared:  0.236
F-statistic:  1110 on 4 and 14352 DF,  p-value: < 2.2e-16
```

Interpretation

Model Overview

- **Dependent Variable (Target)**: spotify_track_popularity
- **Independent Variables (Predictors)**: loudness, danceability, valence, energy

The model specification is:

> **Popularity = 47.410 + 1.863 × Loudness + 40.942 × Danceability - 32.408 × Valence + 5.979 × Energy**

This multiple linear regression model seeks to explain how a combination of song characteristics impacts track popularity.

Interpretation of Coefficients

Term	Estimate	Interpretation
Intercept	47.410	When all predictors are 0, the expected popularity is approximately 47.41.
Loudness	1.863	Each 1 dB increase in loudness is associated with an approximate 1.86 point increase in popularity, holding other variables constant.
Danceability	40.942	Each unit increase in danceability (on its normalized scale) is associated with an approximate 40.94 point increase in popularity, holding other variables constant.
Valence	-32.408	Each unit increase in valence is associated with an approximate 32.41 point decrease in popularity, holding other variables constant.
Energy	5.979	Each unit increase in energy is associated with an approximate 5.98 point increase in popularity, holding other variables constant.

Statistical Significance

All predictor variables are statistically significant with p-values less than 2e-16.
This indicates strong evidence against the null hypothesis, confirming that each predictor meaningfully contributes to explaining variability in track popularity.

Model Fit Summary

Metric	Value	Interpretation
Multiple R-squared	0.2362	Approximately 23.62% of the variance in popularity is explained by the model.
Adjusted R-squared	0.236	After adjusting for the number of predictors, approximately 23.6% of the variance is still explained.
Residual Standard Error	19.84	On average, the model's prediction deviates from the actual popularity by ±19.84 points.

F-statistic	1110 (p-value < 2.2e-16)	The overall model is highly statistically significant, suggesting that at least one predictor is significantly associated with popularity.

Model Evaluation

Compared to the simple linear model, the multiple linear regression model shows an improved explanatory power (R-squared increased from 0.1284 to 0.2362).
Including multiple predictors such as danceability, valence, and energy along with loudness provides a more comprehensive understanding of the factors influencing popularity.

However, while the model fit has improved, a substantial portion of the variance (approximately 76%) remains unexplained.
This indicates that other factors not included in the model (e.g., lyrical content, marketing, artist reputation, release timing) also significantly impact track popularity.

Conclusion

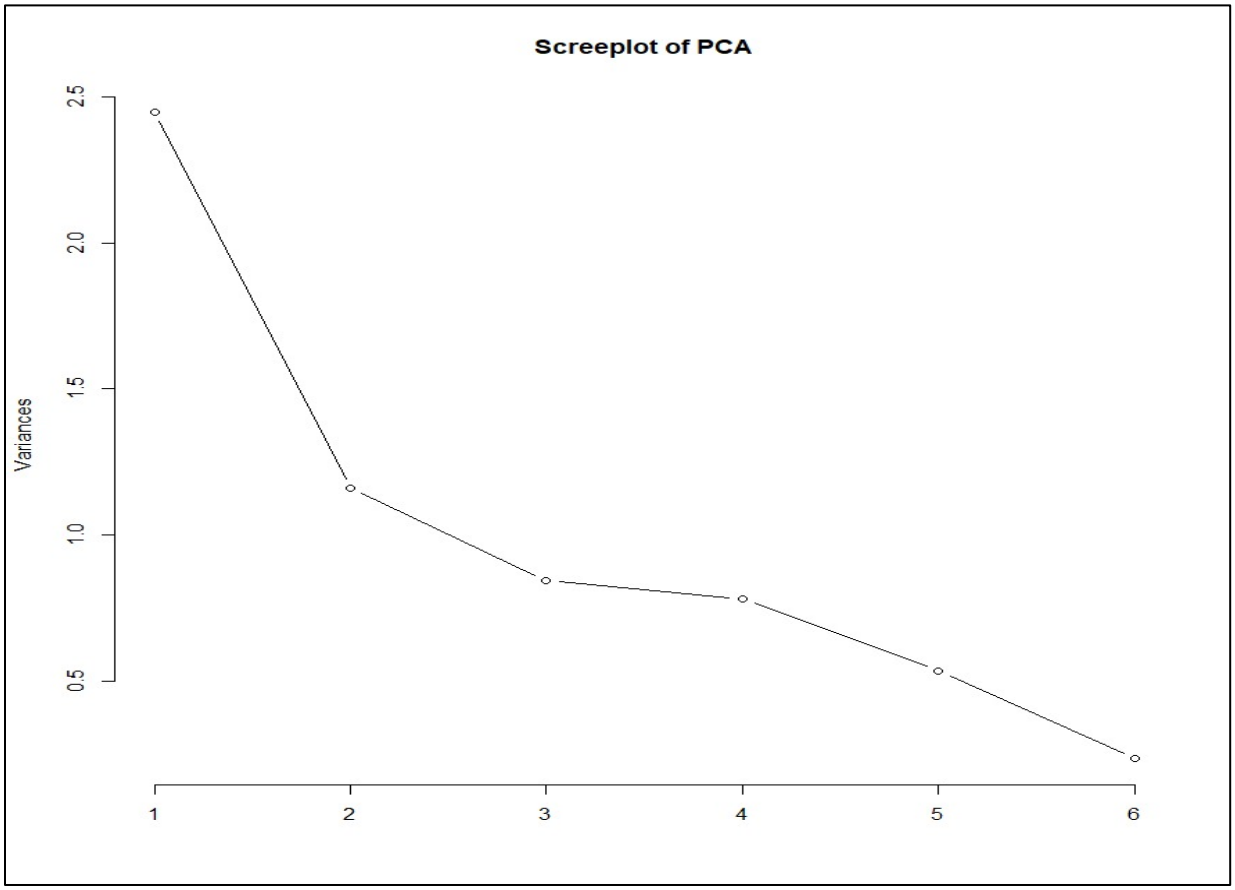
Popularity is positively influenced by higher loudness, greater danceability, and higher energy, whereas higher valence is negatively associated with popularity.
Overall, using a combination of multiple features provides a better predictive model compared to relying on loudness alone.

6. Principal Component Analysis (PCA)

- **Objective:** Reduce dimensionality while preserving variance.

```
#PCA
 pca <- prcomp(scale(num_vars))
 print(summary(pca))
 screeplot(pca, type = "lines", main = "Screeplot of PCA")
```

```
> pca <- prcomp(scale(num_vars))
>  print(summary(pca))
Importance of components:
                          PC1    PC2    PC3    PC4     PC5     PC6
Standard deviation     1.5650 1.0772 0.9179 0.8833 0.73110 0.48296
Proportion of Variance 0.4082 0.1934 0.1404 0.1300 0.08908 0.03888
Cumulative Proportion  0.4082 0.6016 0.7420 0.8720 0.96112 1.00000
>  screeplot(pca, type = "lines", main = "Screeplot of PCA")
> fa_data <- scale(num_vars)
>  psych::scree(fa_data, main = "Scree Plot for Factor Analysis")
>  fa_res <- psych::fa(fa_data, nfactors = 2, rotate = "varimax")
```



Principal Component Analysis (PCA) Interpretation

Principal Component Analysis (PCA) was conducted on the scaled numerical variables to reduce dimensionality while preserving as much variability as possible.
Scaling ensures that variables are normalized, giving each feature equal weight in the analysis.

The following summarizes the PCA output:

	PC1	**PC2**	**PC3**	**PC4**	**PC5**	**PC6**
Standard Deviation	1.5650	1.0772	0.9179	0.8833	0.7311	0.4830
Proportion of Variance	0.4082	0.1934	0.1404	0.1300	0.0891	0.0389
Cumulative Proportion	0.4082	0.6016	0.7420	0.8720	0.9611	1.0000

Detailed Interpretation

1. Standard Deviation of Each Principal Component

- PC1 has the highest standard deviation (1.5650), indicating it captures the most variation in the data.
- PC6 has the lowest standard deviation (0.4830), suggesting it captures the least variation.
- A higher standard deviation implies a principal component explains more variance in the dataset.

2. Proportion of Variance Explained

- PC1 alone explains approximately 40.82% of the total variance.
- PC2 explains an additional 19.34%, while PC3 accounts for a further 14.04%.
- The variance explained decreases progressively from PC1 to PC6.

3. Cumulative Proportion of Variance

- The first two components (PC1 and PC2) together explain about 60.16% of the total variance.
- By including PC3, 74.20% of the variance is explained.
- Adding PC4 increases the cumulative variance explained to 87.20%.
- By PC5, approximately 96.11% of the total variance is captured.
- PC6 brings the cumulative variance explanation up to 100%**.**

Key Observations

- The first four principal components explain approximately 87% of the total variance in the dataset.
- Principal Component 1 (PC1) captures the highest variance, accounting for about 40.8% of the total variance.
- Dimensionality reduction is possible: the original six variables can be reduced to four principal components with minimal information loss.
- PC2, PC3, and PC4 capture smaller, yet significant, patterns in the data structure.
- PC5 and PC6 capture very little variance and mainly represent noise, thus are less important for further analysis.

Conclusion

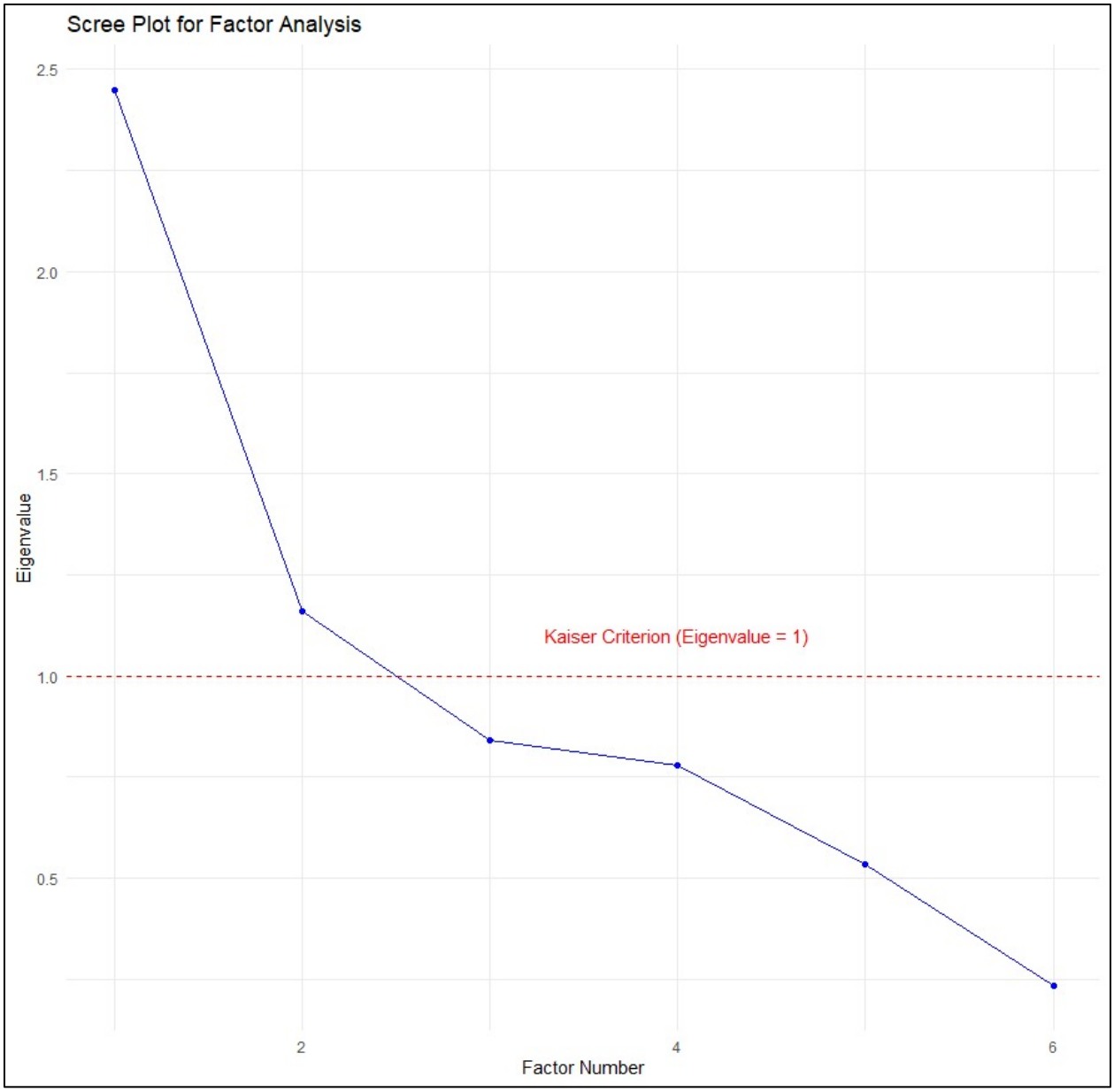
- PC1 is the major source of variation across the dataset.
- PC2, PC3, and PC4 capture smaller but still significant patterns.
- PC5 and PC6 capture very minimal variance and can be considered as noise.
- Reducing the dataset to the first four principal components provides a compact yet highly informative representation of the original data, which is beneficial for further analysis and modeling.

7. Factor Analysis

Factor Analysis is a statistical method used to identify underlying relationships between observed variables. It reduces data complexity by grouping correlated variables into fewer, unobserved factors. This technique helps uncover hidden patterns and is widely used in fields like psychology and marketing for data simplification and insight generation.

```
#Factor analysis
library(psych)
 cor(num_vars)
 fa.parallel(num_vars, fa = "fa")
 fa_result <- fa(num_vars, nfactors = 2, rotate = "varimax")
 print(fa_result)

 cor_matrix <- cor(num_vars)
 eigen_values <- eigen(cor_matrix)$values

 scree_data <- data.frame(
   Factor = 1:length(eigen_values),
   Eigenvalue = eigen_values
 )

 ggplot(scree_data, aes(x = Factor, y = Eigenvalue)) +
   geom_line(color = "blue") +
   geom_point(color = "blue") +
   geom_hline(yintercept = 1, linetype = "dashed", color = "red") +
   annotate("text", x = 4, y = 1.1, label = "Kaiser Criterion (Eigenvalue = 1)", color = "red") +
   labs(
     title = "Scree Plot for Factor Analysis",
     x = "Factor Number",
     y = "Eigenvalue"
   ) +
   theme_minimal()
```

```
>  # View the results
>  print(fa_result)
Factor Analysis using method =  minres
Call: fa(r = num_vars, nfactors = 2, rotate = "varimax")
Standardized loadings (pattern matrix) based upon correlation matrix
                        MR1   MR2    h2      u2 com
spotify_track_popularity  0.29  0.40 0.240  0.7595 1.8
danceability              0.22  0.41 0.213  0.7866 1.5
energy                    1.00 -0.09 1.002 -0.0018 1.0
loudness                  0.68  0.13 0.484  0.5155 1.1
acousticness             -0.61 -0.38 0.515  0.4853 1.7
liveness                  0.09 -0.30 0.097  0.9027 1.2

                       MR1  MR2
SS loadings           1.97 0.58
Proportion Var        0.33 0.10
Cumulative Var        0.33 0.43
Proportion Explained  0.77 0.23
Cumulative Proportion 0.77 1.00

Mean item complexity =  1.4
Test of the hypothesis that 2 factors are sufficient.

df null model =  15  with the objective function =  1.46 with Chi Square =  20914.18
df of  the model are 4  and the objective function was  0.11

The root mean square of the residuals (RMSR) is  0.04
The df corrected root mean square of the residuals is  0.08

The harmonic n.obs is  14357 with the empirical chi square  794.76  with prob <  1e-170
The total n.obs was  14357  with Likelihood Chi Square =  1616.51  with prob <  0

Tucker Lewis Index of factoring reliability =  0.711
RMSEA index =  0.168  and the 90 % confidence intervals are  0.161 0.174
BIC =  1578.23
Fit based upon off diagonal values = 0.98
```



Factor Analysis Interpretation

- Factor Analysis was conducted using two factors, with **varimax rotation** for easier interpretation, and the **minimum residual (minres)** method for extraction.

- **Standardized Loadings (Pattern Matrix)** analysis reveals:

    - **Energy** has a very strong loading (1.00) on **Factor 1 (MR1)**, indicating that it heavily defines this factor.
    - **Loudness** also loads significantly (0.68) on MR1.

- **Acousticness** loads negatively on both MR1 (-0.61) and MR2 (-0.38), showing moderate negative relationships.
- **Spotify Track Popularity** (0.40) and **Danceability** (0.41) load more strongly on **Factor 2 (MR2)**.
- **Liveness** has a relatively low loading overall, with -0.30 on MR2.

- **Interpretation of Factors**:

  - **MR1** appears to represent an **"Energy and Loudness"** dimension.
  - **MR2** seems to capture **"Popularity and Danceability"** characteristics.

- **Communality (h^2) Analysis**:

  - **Energy** has very high communality ($h^2 = 1.002$), meaning it is almost fully explained by the two factors.
  - **Acousticness** and **Loudness** also have reasonably good communalities (~0.48–0.51).
  - **Liveness** shows very low communality ($h^2 = 0.097$), indicating that it is poorly explained by the factors and may not fit well in this model.

- **Uniqueness (u^2) Observations**:

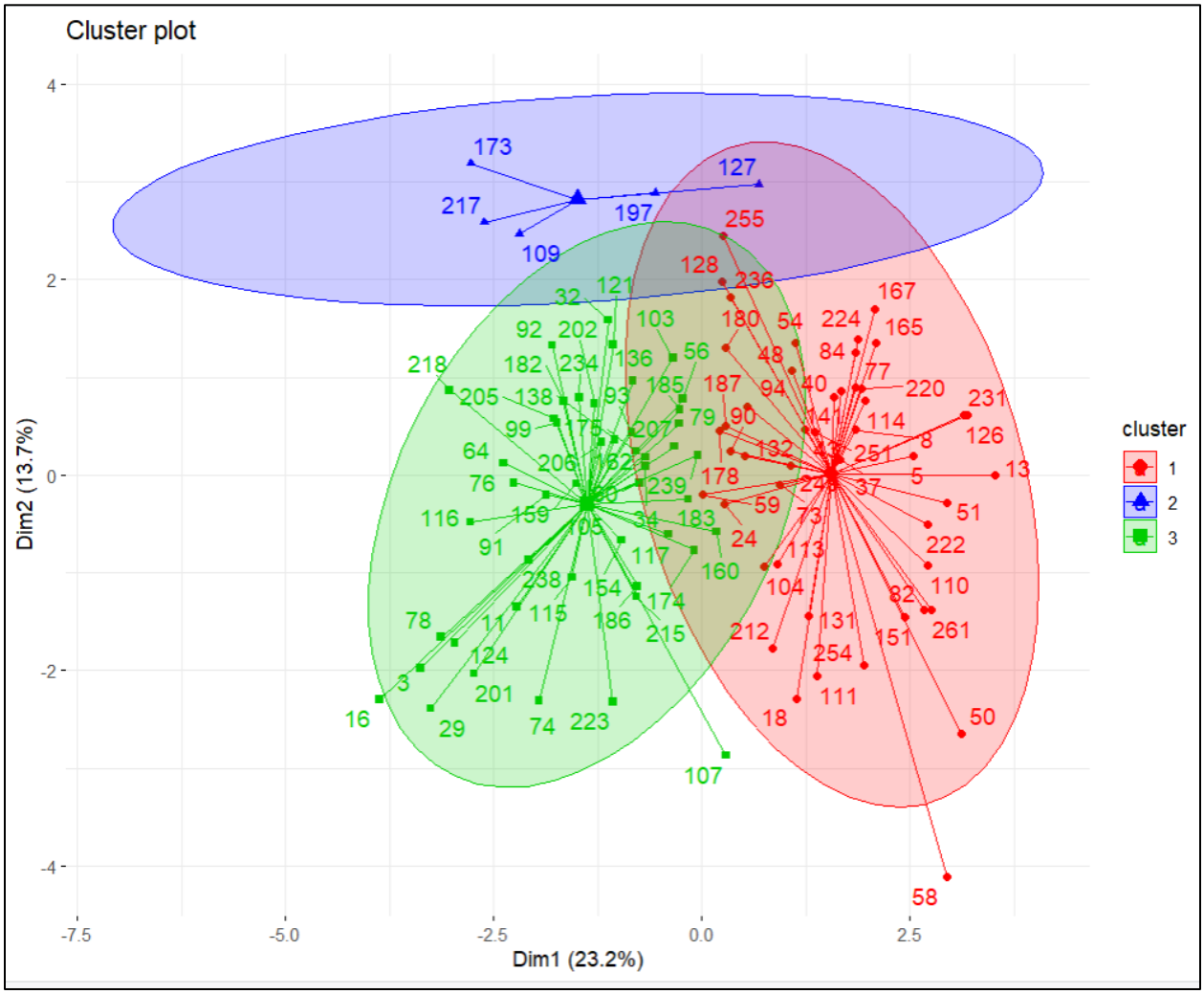
  - **Liveness** has very high uniqueness ($u^2 = 0.90$), suggesting that 90% of its variance is unexplained by the extracted factors.
  - Variables with high uniqueness (>0.5) like Liveness are not well represented by the current factor solution.

- **Variance Explained**:

  - **Factor 1 (MR1)** explains approximately **33%** of the total variance.
  - **Factor 2 (MR2)** explains an additional **10%**.
  - Together, the two factors explain approximately **43%** of the total variance in the dataset.
  - While this level of explained variance is moderate (ideally >50%), it is acceptable for exploratory purposes.

8. Cluster Analysis

- **Objective:** Group similar songs.



Cluster 1	**Red**	Songs that are similar: Maybe **high energy**, **fast tempo**, **loud** (depending on feature values).
Cluster 2	**Blue**	Songs that are **very different** or **special**: Possibly **calm**, **slow**, **unique** songs (because there are few points here).

Cluster 3	**Green**	Songs that are **moderate**: Neither very high nor very low. A **mix** of characteristics.

Plot explanation

Red Cluster:

- Tightly packed → Songs are very similar to each other means high intra-cluster
- Located mainly on one side of the plot (Dim1/DIM2).

Green Cluster:

- More spread out vertically → Songs are somewhat similar but with some variation.

Blue Cluster:

- Very few points, and quite far from other clusters → These songs are **very different** compared to others .

Ellipses:

- The **ellipses (circles)** show the area covered by each cluster.
- **Tight small ellipse** → Songs are **very similar**.
- **Big wide ellipse** → Songs are **more varied**.

Axis: Dim1 and Dim2

- They reduce many features into 2 dimensions to visualize it easily.

So, we **reduce** all the features into **2 new dimensions**:

- **Dim1** (Dimension 1)
- **Dim2** (Dimension 2)

These two dimensions **capture the most important information** (the biggest differences between songs).

- Dim1 explains 23.2% of variance.
- Dim2 explains 13.7% of variance.
- Together they show about 36.9% of the real difference between songs.

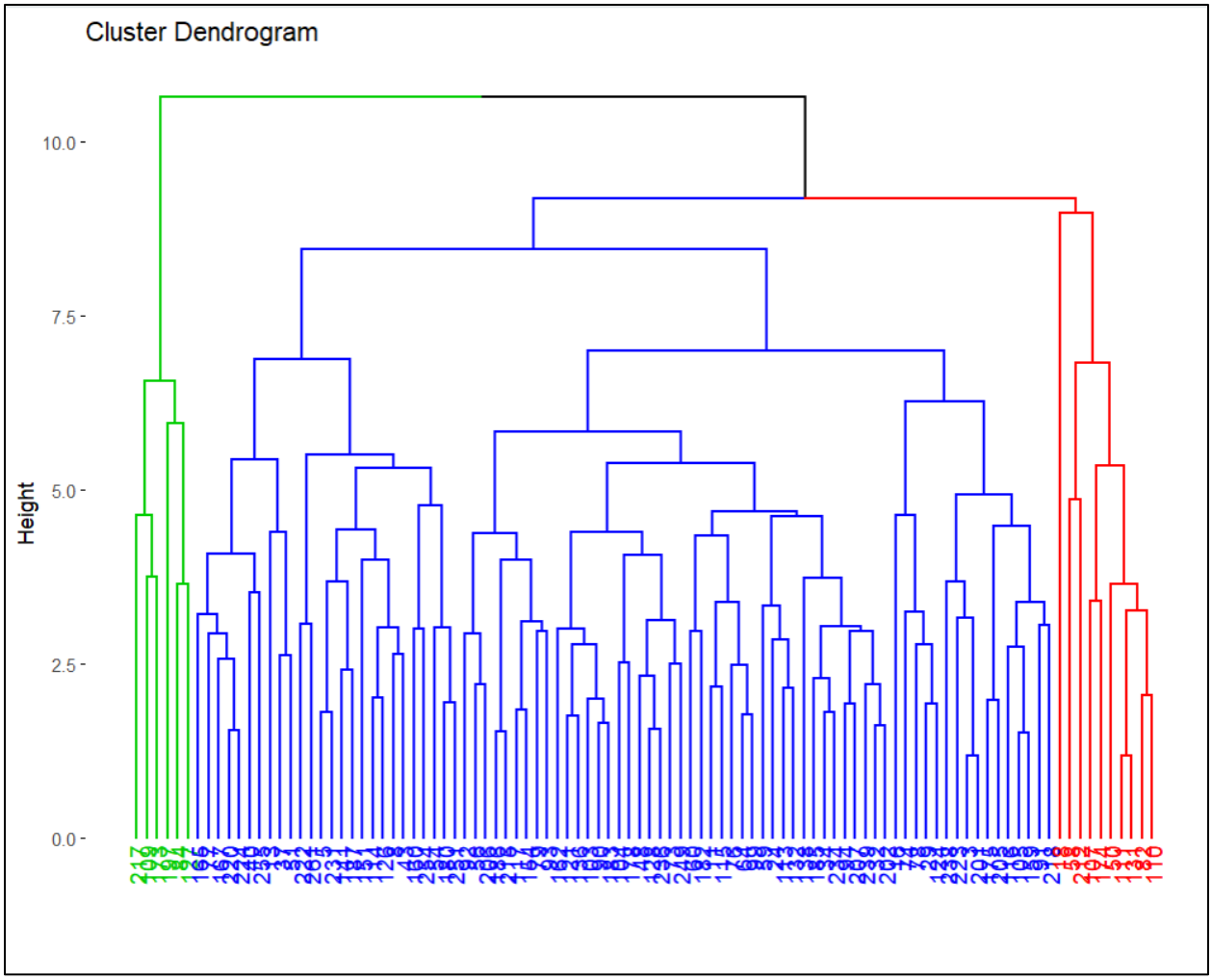
Cluster Dendrogram with their methods

Cluster Dendrogram for the whole Data sheet:

R Code:

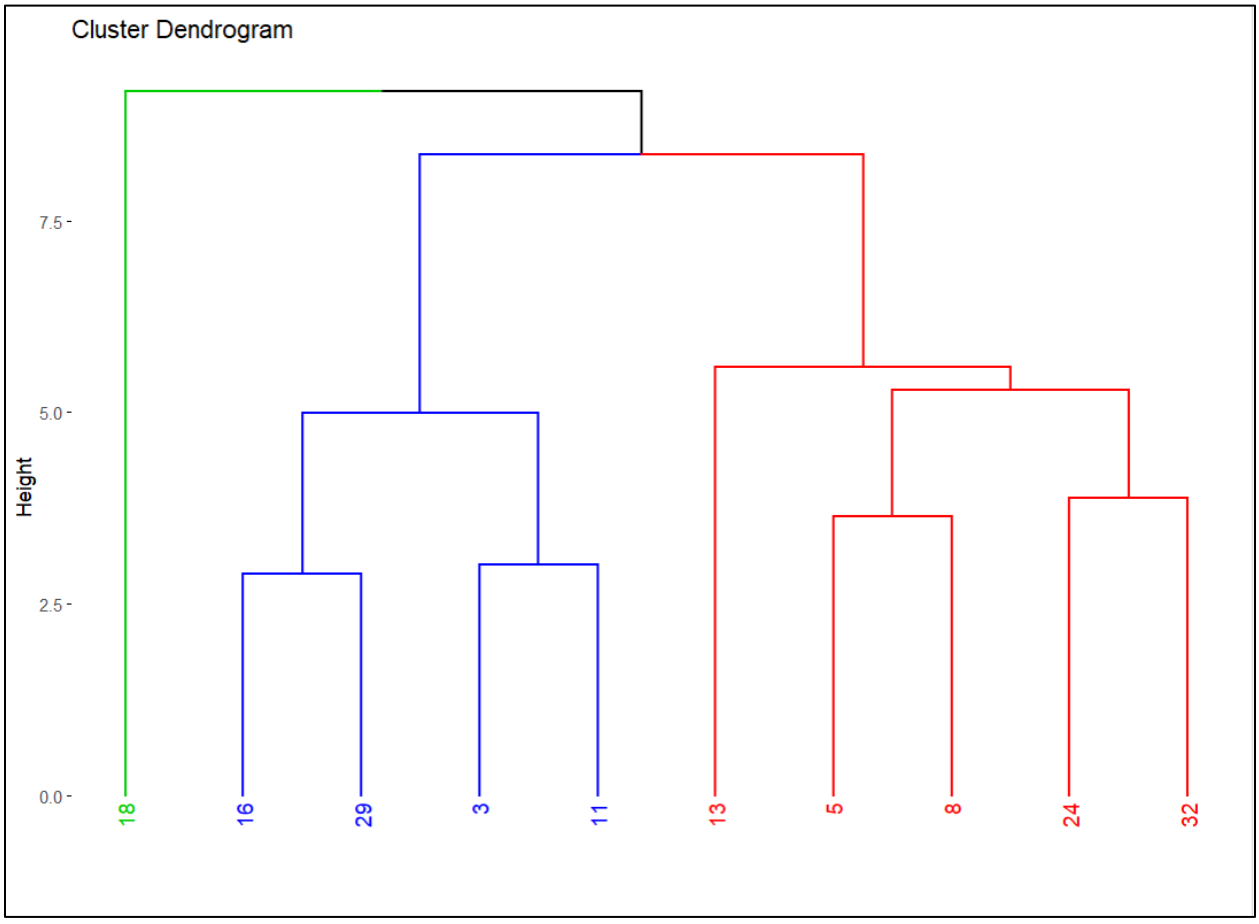
```
plot(res.hc)

fviz_dend(x = res.hc, cex = 0.8, lwd = 0.8, k = 3,

        k_colors = c("green3", "blue", "red"))
```

Output:



Cluster Dendrogram for the top 10 Data rows:

Output:



This dendrogram is the result of hierarchical clustering.
It tells you how similar or different your 10 observations (16, 29, 3, 11, 5, 8, 24, 32, 13, 18) are.
Group similar observations into clusters.
At the bottom, you see:
16, 29, 3, 11, 5, 8, 24, 32, 13, 18.
These are the IDs or row numbers of your dataset.
Each number represents **one data point**

Clustering Interpretation

- **First merges (low height = very similar):**
  - 16 and 29 are **merged early** — they are **very close** (height ~2).
  - 3 and 11 are also **merged early** (height ~3).
  - 24 and 32 merge together around height ~3.5.
  - 5 and 8 are also close.

- **Second level merges:**
  - (16 and 29) cluster then merges with (3 and 11) — around height ~5.
  - (5 and 8) then merge with (24 and 32) — slightly higher than ~5.

- **Higher level merges (less similar):**
  - Cluster (5,8,24,32) merges with 13 — meaning 13 is slightly more different but still relatively close.
  - Everything finally merges with 18 at the very **top height (~9)** — meaning **observation 18 is the most different** from all others.

Height Axis (Left side)

The **Height** is **Euclidean distance** between points or clusters.

- **Low height (~2-3)** = very **similar** observations.
- **High height (~9)** = **very different** observations.

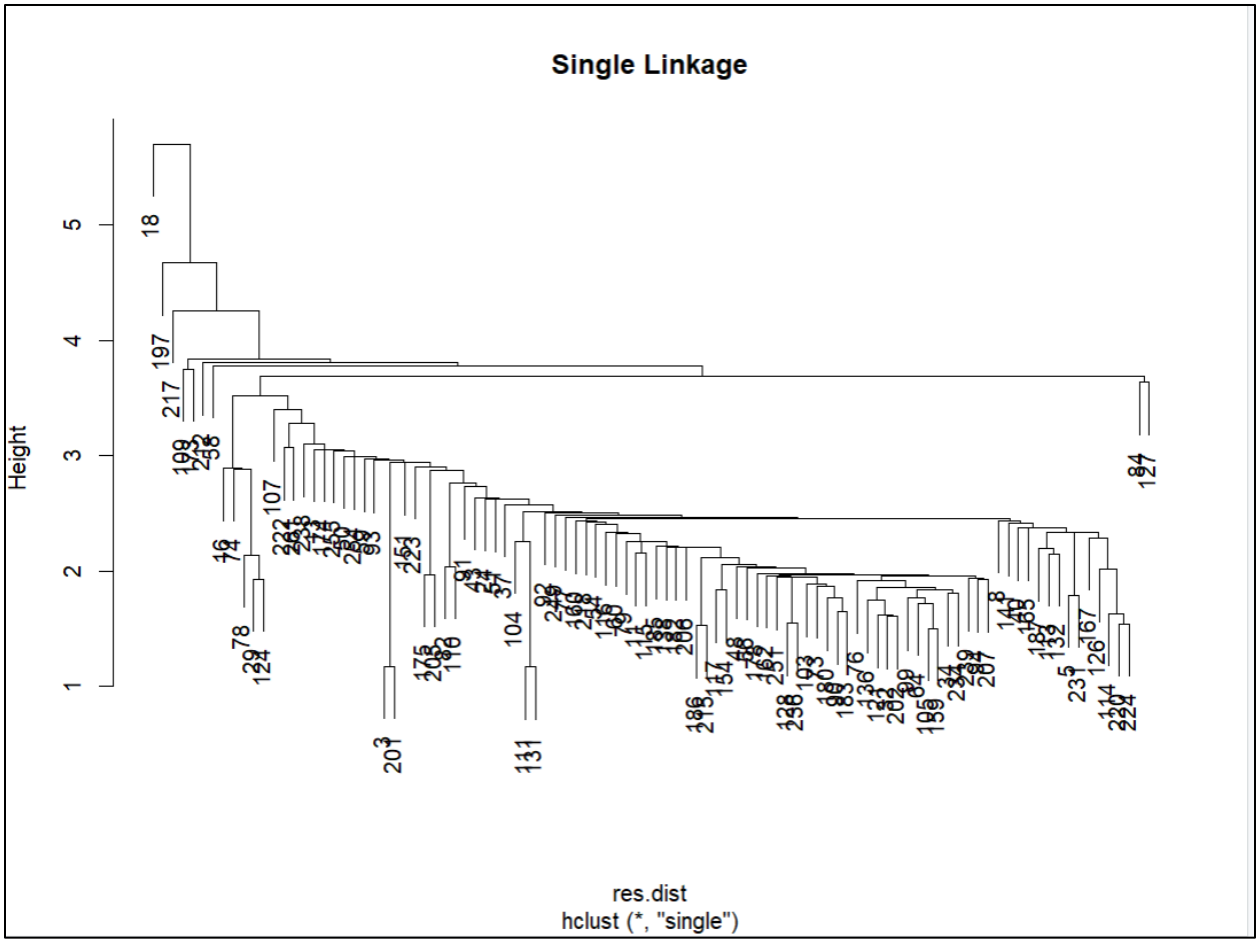
Single Linkage Method

Cluster Dendrogram for the whole Data sheet:

R Code:

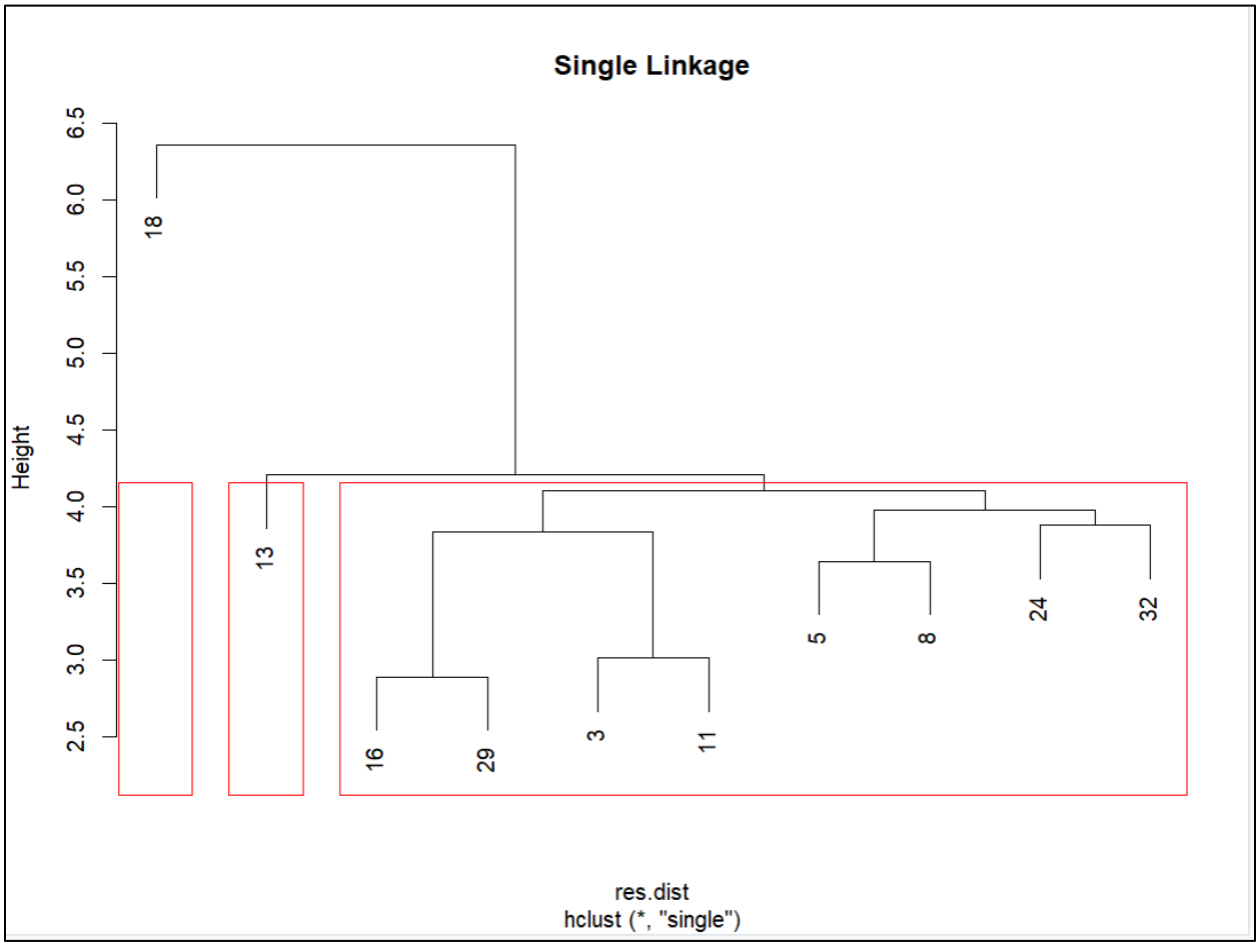
```
# Single linkage

res.hc_single <- hclust(d = res.dist, method = "single")

# Plot dendrogram for single linkage

plot(res.hc_single, main = "Single Linkage")
```

Output:


Single Linkage
Height
1
2
3
4
5
res.dist
hclust (*, "single")

Cluster Dendrogram for the top 10 Data rows:

Output:



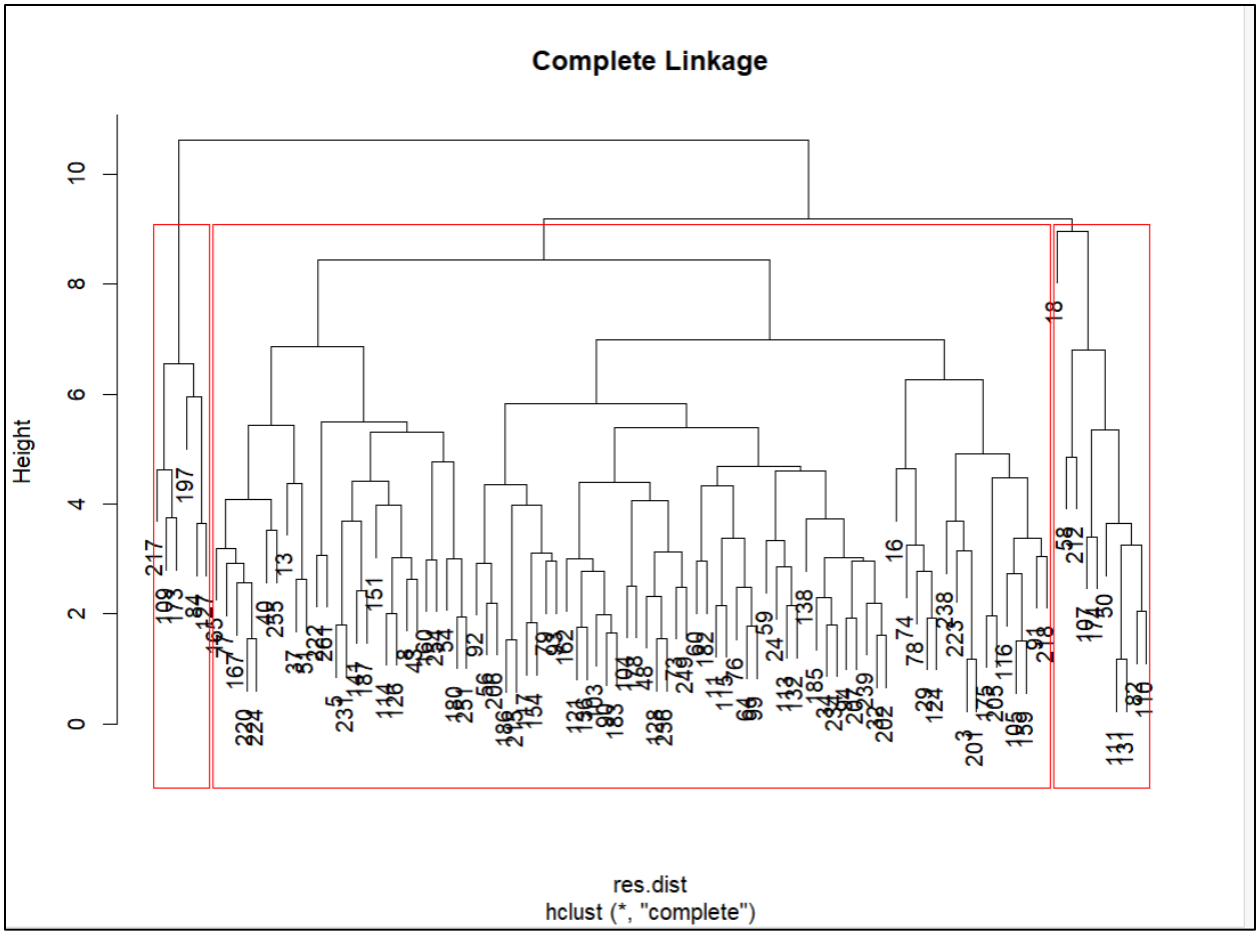
Complete Linkage Method

Cluster Dendrogram for the whole Data sheet:

R Code:

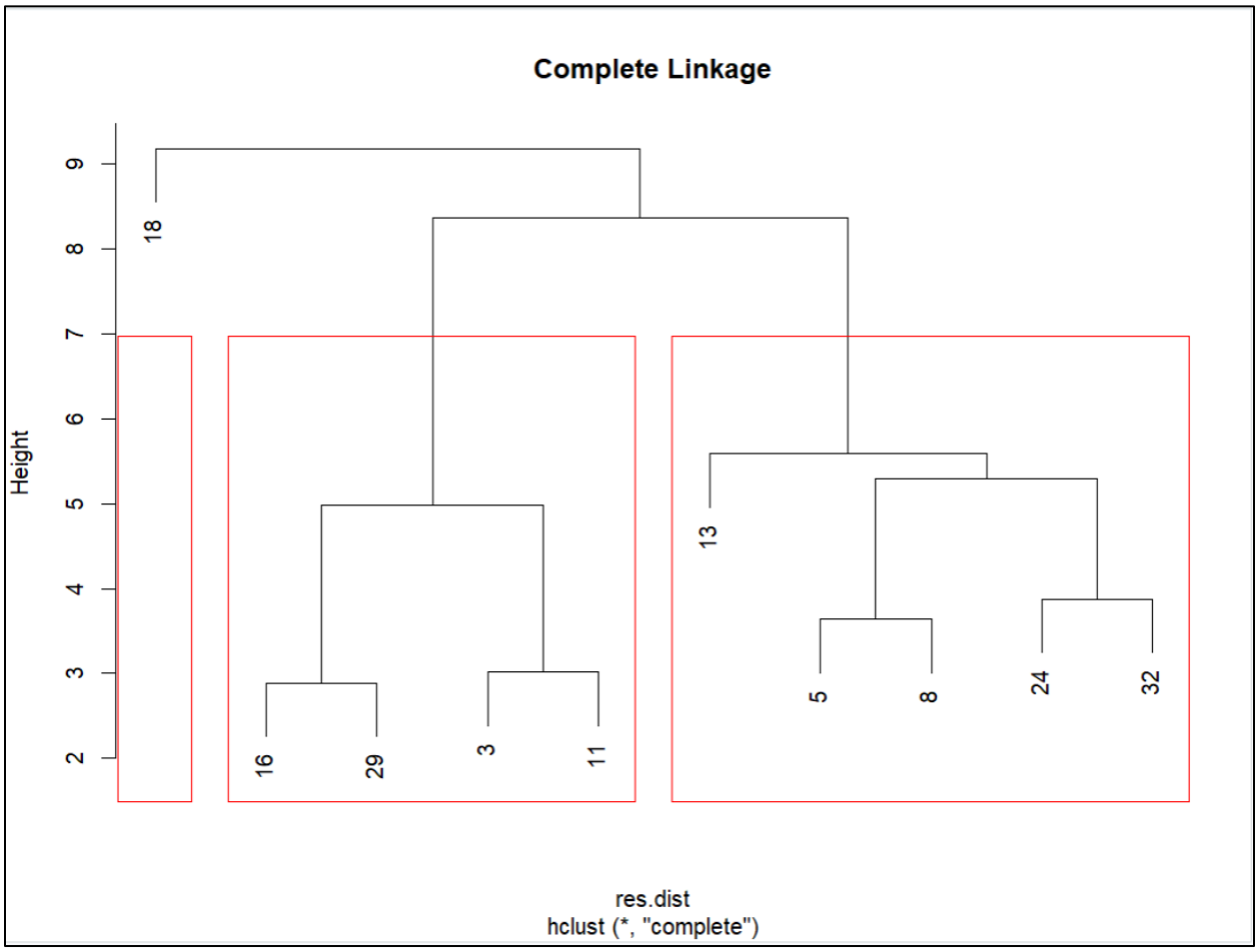
```
res.hc_complete <- hclust(d = res.dist, method = "complete")

# Plot dendrogram for complete linkage

plot(res.hc_complete, main = "Complete Linkage")
```

Output:


Complete Linkage
Height
0
2
4
6
8
10
res.dist
hclust (*, "complete")

Cluster Dendrogram for the top 10 Data rows:

Output:



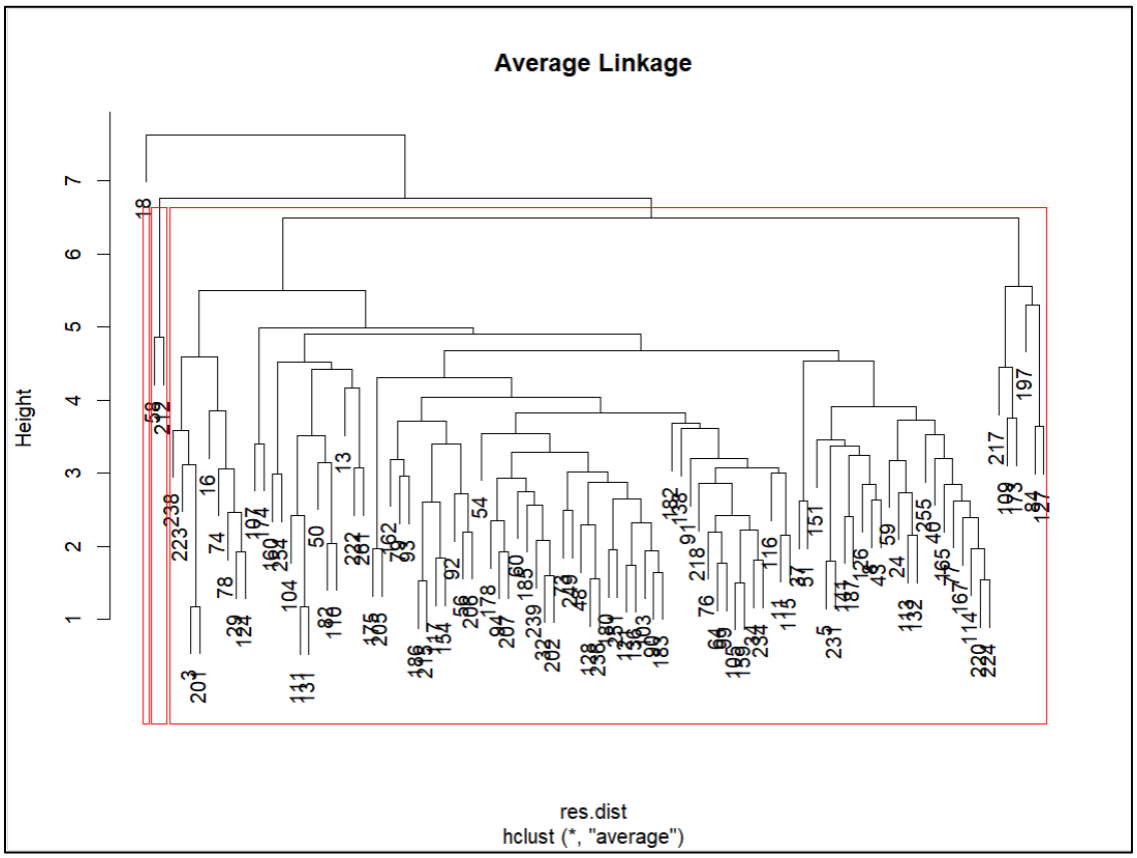
Average linkage Method

Cluster Dendrogram for the whole Data sheet:

R Code:

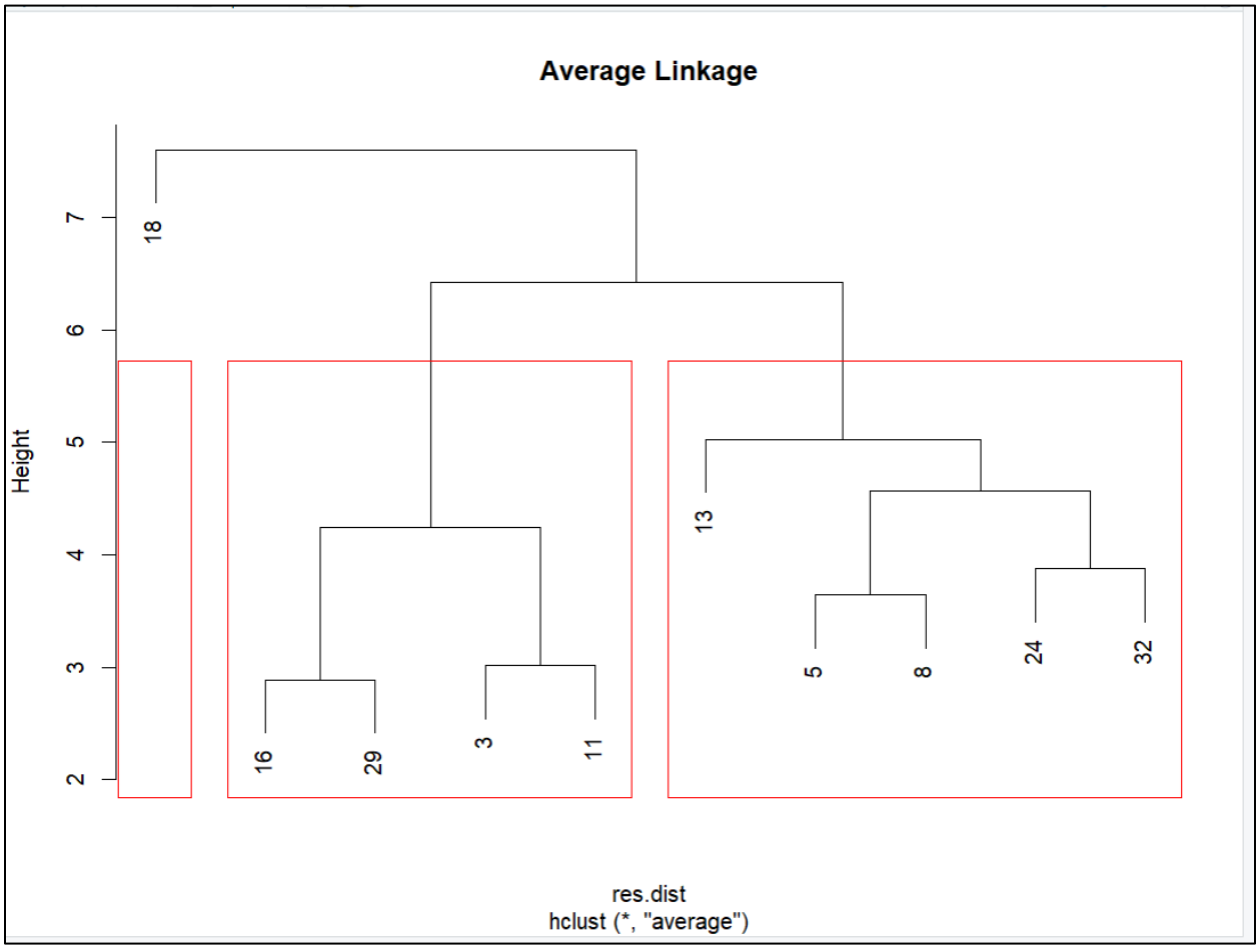
```
res.hc_average <- hclust(d = res.dist, method = "average")

# Plot dendrogram for average linkage

plot(res.hc_average, main = "Average Linkage")
```

Output:



Cluster Dendrogram for the top 10 Data rows:

Output:



9. Discussion

In this study, a combination of data visualization, regression modeling, Principal Component Analysis (PCA), and clustering techniques was used to explore and understand the relationships between various audio features of songs and their popularity.

1. Data Visualization Insights

Data visualizations played a crucial role in the initial stages of the analysis. Through histograms, boxplots, and scatterplots, key patterns and distributions among song features were identified. For example, features like **energy**, **danceability**, and **loudness** showed wide variation across songs, while **acousticness** and **liveness** appeared to have more skewed distributions. Visualization helped in detecting outliers, understanding feature behavior, and guiding the selection of appropriate modeling techniques.

2. Regression Modeling Findings

Regression models provided important insights into the predictors of song popularity:

- **Multiple linear regression** highlighted that **energy** was a strong positive predictor of track popularity. Songs with higher energy scores tended to be more popular.
- Other features like **danceability** and **loudness** also showed moderate associations with popularity.
- However, **multicollinearity** among predictors was observed, particularly between energy and loudness, which required careful interpretation of coefficients and model performance.

The regression models emphasized the importance of energetic and lively audio characteristics in influencing the success of a song.

3. Dimensionality Reduction with PCA

Principal Component Analysis (PCA) was applied to reduce the dimensionality of the dataset while retaining most of the variability:

- The first four principal components captured approximately **87%** of the total variance, making it possible to reduce from six original features to four principal components without substantial information loss.
- PCA revealed that features related to **energy** and **loudness** were dominant in the first principal component, while **popularity** and **danceability** contributed more to the second component.

- The application of PCA streamlined subsequent analyses, particularly clustering, by working with uncorrelated dimensions.

4. Clustering of Songs

Clustering techniques grouped songs based on similarities in their audio features:

- Using the reduced feature set from PCA, songs were clustered into distinct groups that shared similar musical characteristics.

- Clear groupings were observed between songs with high energy and loudness, and those that were softer and more acoustic.

- Clustering validated the insights from regression and PCA, strengthening the interpretation that certain audio profiles are more characteristic of popular songs.

5. Challenges Encountered

Several challenges were encountered during the analysis:

- **Class imbalance**: There was an uneven distribution between popular and non-popular songs. Popular songs constituted a smaller proportion of the dataset, which may have biased model performance and required careful evaluation of metrics beyond simple accuracy.

- **Multicollinearity**: Strong correlations between variables, particularly between **energy** and **loudness**, affected multiple regression models. This multicollinearity made it difficult to isolate the individual impact of each predictor without relying on dimensionality reduction techniques like PCA.

10. Conclusion

This study comprehensively analyzed the relationships between various audio features and song popularity using a combination of data visualization, statistical modeling, dimensionality reduction, and clustering techniques.

Initially, data visualization provided valuable insights into the distribution and behavior of the features. Patterns such as the positive skewness of acoustic ness and the wide variation of energy levels across songs guided the focus of the analysis. These preliminary observations were critical for selecting appropriate models and techniques for deeper investigation.

Multiple linear regression revealed that **energy** was the most significant predictor of song popularity, with **loudness** and **danceability** also contributing moderately. The selection of **loudness** as a key predictor was further justified through multivariate visualization, where it demonstrated a strong correlation with popularity. However, multicollinearity between energy and loudness was a notable challenge, necessitating careful interpretation of the model outputs.

To address dimensionality and correlation issues, Principal Component Analysis (PCA) was performed. PCA successfully reduced the six original features into four principal components while retaining approximately **87%** of the total variance. This reduction not only simplified the dataset but also enhanced the interpretability and efficiency of subsequent analyses, especially clustering.

Factor Analysis further supported the findings by identifying two main underlying factors: one dominated by **energy** and **loudness**, and the other associated with **popularity** and **danceability**. This confirmed the multi-dimensional nature of the data, where different clusters of features contributed differently to song success.

Clustering techniques, applied after PCA, grouped songs into distinct categories based on their audio characteristics. Songs with high energy and loudness clustered together, emphasizing the trend that more energetic tracks tend to gain greater popularity.

Throughout the process, challenges such as **class imbalance** between popular and non-popular songs and **multicollinearity** among features were encountered. These were addressed through careful model selection, feature scaling, dimensionality reduction, and critical evaluation of the results.

In conclusion, the analysis demonstrated that songs exhibiting higher energy and loudness levels are more likely to achieve higher popularity scores. Furthermore, dimensionality reduction and clustering revealed meaningful groupings within the data, offering valuable insights for future work such as music recommendation systems or predicting song success based on audio characteristics. Despite the challenges, the combination of statistical and machine learning techniques provided a robust framework for understanding the complex interplay between song features and popularity.

11. References

- **Dataset: Spotify song Audio Features (source citation as per your dataset)**
- **R Packages:**

  library(ggplot2)

  library(dplyr)

  library(ggcorrplot)

  library(corrplot)

  library(psych)

  library(factoextra)