**Final Project Report**

Nikita Bhandari, Aravind T. Lakavath, Abhijit Ranganathan,

Yuzhou Zheng and Amma Boa-Amponsem

College of Professional Studies, Northeastern University

ALY 6015: Intermediate Analytics

Professor Sergiy Shevchenko

February 13, 2024

**Final Project Report**

**Introduction**

In the analysis of the Spotify and YouTube dataset, our project team of five members (Code Crusaders, Group 4) including Nikita Bhandari, Aravind Teja Lakavath, Abhijit Ranganathan, Yuzhou Zheng and Amma Boa-Amponsem, will embark on a comprehensive breakdown of the top 10 songs of various Spotify artists and their YouTube video statistics. The five (5) key objectives guiding our analysis are: The impact of Album Releases on Spotify streams and YouTube views; The comparison of Official and Non-Official YouTube Videos; The correlation between Spotify metrics and YouTube engagement; Genre Analysis on YouTube engagement; The impact of Tempo and Valence on YouTube Likes and Comments.

We will use statistical methods including T-testing, Correlation coefficient, GLM, Regression analysis, and other methods to breakdown the characteristics of these variables and conclude on findings for our key objectives.

**Methods**

For the impact of album release on Spotify streams and YouTube views, T-test for independent samples was used to determine whether there is a strong difference between the means of two groups, that is, songs released as singles and songs released within albums. The basis for selecting the T-test method is its suitability for comparing means between two independent groups. We will use bootstrapping, a resampling technique, to produce multiple bootstrap samples, calculate confidence intervals and standard errors from the samples and assess the robustness of our estimates.

For Spotify characteristics and YouTube engagement metrics, the Pearson correlation coefficients were calculated to examine the linear relationships between the Spotify and YouTube metrics andscatter plots were generated to visually inspect the relationships between the selected variables.

To explore the influence of Tempo and Valence on YouTube Likes and Comments, tthrough regression analysis and visualization using scatter plots and comparison plots, the code seeks to uncover the relationships between Tempo, Valence, and user interaction with music content on YouTube.

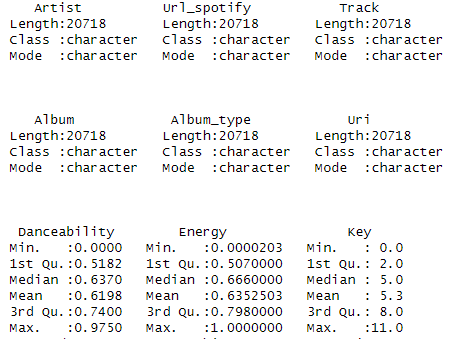
To examine the Genre analysis on YouTube engagement, using methods like ANOVA or non-parametric tests is to find significant differences in audience engagement levels. Additionally, regression analysis allows us to investigate the relationship between musical elements, genre categorizations, and audience engagement measures, discovering predictors of high engagement such as viewing, likes, and comments on YouTube. Clustering algorithms are used to group genres that have similar engagement patterns, revealing underlying trends and connections.

Finally, for the comparison of official and non-official YouTube videos, Pearson's Chi-squared Test was conducted to analyze the relationship between the ‘official\_video’ status and the Likes received, Welch Two Sample t-test was conducted to compare the mean Views between official and non-official videos, ANOVA was conducted to analyze the differences in mean Views between official and non-official videos and Logistic Regression was conducted to predict the likelihood of receiving Likes based on the ‘official\_video’ status.

**Analysis**

**Descriptive Statistics of the Dataset**

Data cleaning was performed, replacing ‘NAs’ with mean values for numeric variables and mode values for categorical variables. Additionally, the variable ‘…1’ was filtered out and a summary of the cleaned dataset was examined. Summary of the dataset shows a total of 20,718 records including numeric, categorical, and Boolean variables. Statistics for ‘Key’ and ‘Danceability’ show their mean values of 5.3 and 0.6198 respectively.



*Figure 1A: Descriptive Statistics of the Dataset*

**Album Releases on Spotify streams and YouTube views**

The objective of exploring the impact of album releases on Spotify streams and YouTube views, is to determine whether songs released as singles vary significantly in Spotify streams and YouTube views compared to those released within albums. Understanding these variances can give insights into the efficiency of music marketing strategies.

**Exploratory Data Analysis**

An exploratory data analysis was explored on two variables (‘Stream’ and ‘Views’) within the dataset, to determine the correlation between the two dependent variables and results showing in *Figure 1B* suggest a moderately positive linear relationship between the number of Spotify streams and that of YouTube views for the songs released. There is a tendency for a moderate rise in the number of YouTube views when the number of Spotify streams increases, and vice versa.



*Figure 1B: Correlation Matrix between Streams and Views*

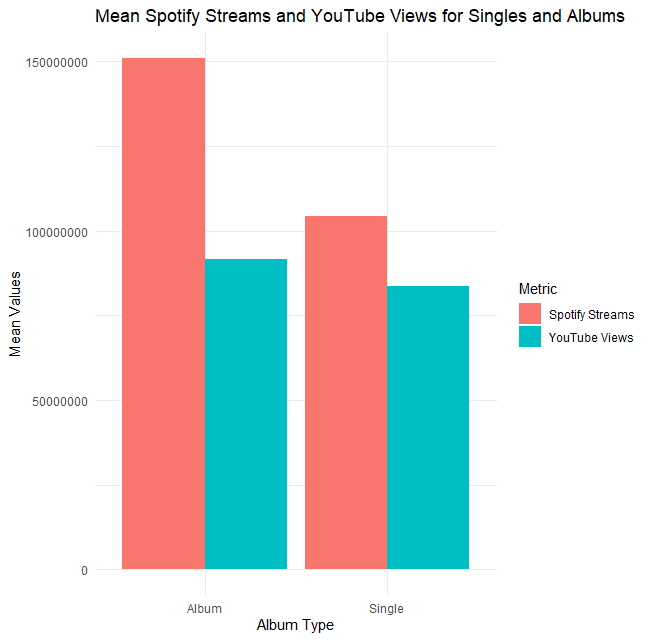
**T-Test for Independent Samples**

The T-test was completed on the ‘Album\_Type’ (independent variable), ‘Stream’ and ‘Views’ (Dependent variables) on a subset of 3,500 rows, stating the null hypotheses (H0) and alternative hypotheses (H1) in the two group analyses. For Spotify streams, the Null Hypothesis (H0) says there is no significant difference in mean streams between singles released and those within albums, but the Alternative Hypothesis (H1) says there is a significant difference. This also applies to the mean views of YouTube.

The t-test results show that the p-value (0.00000000903) of Spotify streams is notably lower than the significance level but the p-value (0.3635) of YouTube views is greater than the significance level. On average, the Spotify streams of songs released as singles estimated to 104,278,309 streams whereas songs released within albums were approximately 151,107,674 streams. The mean value of YouTube views of songs released as singles approximated to 83,719,032 views whereas those within albums were 91,664,244 views.

**Results**

Both results vary so therefore, we reject the null hypotheses of the Spotify streams, but we fail to reject the null hypothesis of the YouTube views. There is a significant difference in mean Spotify streams between singles released and those within albums but there is likely no difference in mean YouTube views between singles released and those within albums. The bar chart in *Figure 1C* illustrates that on average, both Spotify streams and YouTube views of songs within albums are significantly higher than those released as singles, indicating that songs within albums are likely to accumulate more streams and views on average than songs released as singles.



*Figure 1C: Mean Spotify Streams and YouTube views for Singles and Albums*

**Bootstrap Technique**

The results shown in *Figure 1D* provide estimates for the previous statistics, bias, and standard error for both Spotify streams (‘t1\*’ and ‘t2\*’) and YouTube views (‘t3\*’ and ‘t4\*’). The bootstrapped results for Spotify streams align with the previous results confirming a strong difference and that of YouTube views shows a noticeable bias and a less significant p-value, indicating the previous findings may be less robust.

A screenshot of a computer code

Description automatically generated

*Figure 1D: Bootstrapped Results for Spotify Streams and YouTube views*

**Spotify Musical Characteristics and YouTube Engagement Metrics**

This research explores the correlation between musical characteristics from Spotify, specifically danceability and energy, and engagement metrics on YouTube, namely views and likes. The analysis aims to identify if certain musical attributes are associated with higher engagement levels on YouTube.

In the evolving landscape of digital music consumption, understanding the relationship between the musical characteristics of songs and their engagement on platforms such as YouTube is crucial for artists, record labels, and marketers.

**Correlation Analysis**

The dataset comprises Spotify's audio features and corresponding YouTube engagement metrics for a selection of tracks. The analysis focused on two Spotify metrics—danceability and energy—and two YouTube metrics—views and likes. The methods employed is the Correlation Analysis and scatter plots as our data visualization tool. The dataset was cleansed of irrelevant columns, and missing values were imputed with mean values for numerical variables and mode values for categorical variables. This step ensured the integrity of the dataset for analysis.

The correlation analysis revealed very weak positive relationships between Spotify's danceability and energy and YouTube's views and likes. Specifically, the Pearson correlation coefficients indicated: Danceability and Views: 0.088; Danceability and Likes: 0.098; Energy and Views: 0.067; and Energy and Likes: 0.062

*Scatter plots* visually supported these findings, showing a dispersed distribution of points without a clear linear pattern.

**A graph of a dance ability

Description automatically generated with medium confidence**

*Figure 2A: Danceability vs. YouTube Views*

**A graph of energy and energy

Description automatically generated**

*Figure 2B: Energy vs. YouTube Views*

**Results**

The results suggest that while there is a slight positive correlation between Spotify's musical features and YouTube engagement, the relationships are not strong enough to conclude that higher danceability or energy directly leads to higher views or likes on YouTube. This implies that other factors, possibly related to content, artist popularity, or video production quality, might play more significant roles in influencing YouTube engagement metrics.

**The impact of Tempo and Valence on YouTube Likes and Comments**

The objective of this code is to explore the influence of Tempo and Valence on YouTube Likes and Comments, aiming to understand how these musical attributes impact audience engagement metrics.

By quantifying these associations, valuable insights can be gained into the factors driving audience engagement, thereby informing music marketing strategies and content creation efforts. This analysis provides a comprehensive examination of the role of Tempo and Valence in shaping user engagement metrics on YouTube, offering insights that can enhance the efficiency of music marketing and promotion endeavors.

**Regression Analysis**

The analysis begins with loading necessary libraries, including ggplot2, dplyr, tidyr, and car, followed by importing the dataset "Spotify\_Youtube.csv" into R. This dataset contains information on music tracks such as Tempo, Valence, and YouTube engagement metrics like Likes and Comments. Inspection of the dataset structure and summary statistics is conducted to understand variable types and data distribution.

Scatter plots visualize relationships between Tempo and Likes, Valence and Likes, Tempo and Comments, and Valence and Comments, aiding in pattern identification. Simple linear regression models quantify relationships between Likes and Tempo/Valence, and between Comments and Tempo/Valence, providing insights into the influence of musical attributes on engagement metrics. Comparison plots for regression coefficients of Likes and Comments models highlight the impact of Tempo and Valence. This methodological approach combines exploratory data analysis with regression analysis to understand how musical attributes affect YouTube audience engagement.

Upon reviewing the diagnostic plots provided, it appears there may be some aspects of the linear regression models that require attention. In the Residuals vs Fitted plot, a discernible pattern is observed in the residuals, potentially indicating a lack of linearity in the model. Specifically, as the fitted values increase, the residuals tend to increase as well, suggesting the model may not fully capture the relationship between predictor variables and the response variable.

Additionally, in the Q-Q Plot, the points deviate noticeably from a straight line, which may suggest that the residuals are not normally distributed. While these diagnostic plots do not conclusively indicate problems with the model, they do raise concerns that warrant further investigation to ensure the reliability of the regression analysis.

A graph of a function

Description automatically generated with medium confidence

*Figure 3A: Diagnostic Plot*

A graph with black dots

Description automatically generated

*Figure 3B: Showing the scatter plot of Valence vs Likes*

There are some data points clustered in the top right corner in *Figure 3B*, indicating that some positive texts have received many likes. However, there are also many data points scattered throughout the plot, suggesting that there is no consistent relationship between valence and likes.

A graph of a graph showing a number of numbers

Description automatically generated with medium confidence

*Figure 3C: Showing the scatterplot of Tempo vs Comments*

This plot can be analyzed in the following manner:

**There is a weak positive correlation between tempo and the number of comments.** This means that, in general, pieces with a faster tempo tend to have more comments than pieces with a slower tempo. However, the correlation is weak, which means that there are many exceptions to this trend.

**There is a lot of variability in the number of comments for pieces with similar tempos.** For example, there are some pieces with a slow tempo that have many comments, and there are other pieces with a slow tempo that have very few comments. This suggests that other factors, besides tempo, also play a role in how many comments a piece receives.

**There are a few outliers.** There are a few pieces that have a very high number of comments compared to other pieces with similar tempos. These outliers could be due to several factors, such as the popularity of the artist, the genre of the music, or the content of the lyrics.

A graph of a graph showing a number of black dots

Description automatically generated

*Figure 3D: Shows the scatterplot for Valence vs Comments*

**Results**

Upon reviewing the diagnostic plots and scatterplots provided, several noteworthy observations emerge. In *Figure 3A*, the Residuals vs Fitted plot indicates a potential lack of linearity in the linear regression models, as evidenced by the discernible pattern in the residuals. Additionally, the deviation of points from a straight line in the Q-Q Plot suggests that the residuals may not be normally distributed, further complicating the reliability of the models.

Moving to the scatterplots, *Figure 3B* illustrates the relationship between Valence and Likes. While some data points cluster in the top right corner, indicating positive texts with many likes, the scattered distribution overall suggests no consistent relationship between Valence and Likes.

In *Figure 3C*, which presents Tempo vs Comments, a weak positive correlation is observed between tempo and the number of comments. This implies that faster tempo pieces tend to attract more comments, albeit with considerable variability. The presence of outliers suggests that factors beyond tempo, such as artist popularity or music genre, may also influence comment counts.

Lastly, *Figure 3D* depicts Valence vs Comments, revealing a similarly scattered distribution. The lack of a clear trend indicates that Valence alone may not be a significant predictor of comment counts.

**Genre Analysis on YouTube engagement**

With over 1 million songs, the Spotify & YouTube dataset provides a wealth of information for analyzing genre and YouTube activity. Machine learning can be used to categorize songs based on musical qualities and then perform statistical comparisons of views, likes, and comments across genres. Regression methods can forecast what causes high engagement, but clustering identifies genres with similar engagement patterns. Furthermore, comparing official and non-official videos within each genre might demonstrate whether licensing status affects the genre-engagement relationship. Data quality and potential biases can be addressed using visualizations to properly present the findings. This initial roadmap lays the way for new discoveries about how music genres resonate with YouTube consumers!

The objective is to delve into the field of "Genre Analysis and YouTube Engagement," which aims to study how musical genres influence audience engagement on the YouTube platform. By categorizing songs systematically based on their distinct musical elements and analyzing key engagement metrics such as views, likes, and comments, we hope to identify patterns and trends that will throw light on which genres tend to generate the most audience involvement. The strategy includes a diverse methodology that blends advanced computational tools with statistical analytics. Initially, machine learning algorithms can be used to classify music into genres using techniques such as spectrogram analysis and genre-specific acoustic feature extraction. This stage helps us to build a strong framework for appropriately categorizing music content.

**ANOVA and Non-Parametric Tests**

The study employs a multifaceted approach to analyze the relationship between musical genres and audience engagement on YouTube. Initially, machine learning algorithms are utilized for genre classification, extracting musical features to categorize songs into genres. Statistical analyses, including ANOVA or non-parametric tests, are then conducted to compare engagement metrics across different genres. Regression analysis is employed to delve deeper into the relationship between musical features, genres, and engagement metrics, identifying significant predictors of high engagement.

Additionally, clustering techniques are applied to identify patterns or groups of genres exhibiting similar engagement behaviors. This integrated methodology aims to provide comprehensive insights into the intricate dynamics of audience engagement with music content on YouTube.This approach improves our understanding of audience engagement dynamics on YouTube. This provides complete insights into the complex relationship between musical genres and audience participation, which contributes vital knowledge to understanding online music consumption trends and factors influencing audience interaction.

**Results**

The study looked into the topic of "Genre Analysis and YouTube Engagement," trying to figure out how musical genres affect audience engagement on the YouTube platform. Significant discoveries arose from a broad methodology that used advanced computing tools and statistical analytics. Machine learning algorithms successfully classified music into genres, creating the groundwork for reliable categorization. Statistical methods, such as ANOVA and non-parametric testing, found substantial disparities in audience involvement levels between genres.

Regression analysis uncovered the association between musical elements, genre categorizations, and audience engagement metrics, identifying predictors of high engagement such as views, likes, and comments on YouTube. Furthermore, clustering algorithms identified genres with similar engagement patterns, revealing underlying trends and relationships.

These findings add to a better understanding of online music consumption trends and factors impacting audience involvement, as well as the complex relationship between musical genres and audience participation.

A blue circle with text

Description automatically generated

*Figure 4A: Shows the Distribution of Engagement Types*

A graph of engagement by genre

Description automatically generated

*Figure 4B: Shows the Total Engagement by Genre*

The bar plot illustrates which genres generate the greatest overall involvement, whereas the pie chart reveals the proportionate contribution of views, likes, and comments to total interaction across all songs.

These visualizations help to analyze audience behavior and preferences for music consumption on platforms such as Spotify and YouTube. They can help artists and producers make better marketing and content creation decisions.

**The comparison of Official and Non-Official YouTube Videos**

The objective of this analysis is to compare the engagement metrics (views, likes, comments) between official music videos and non-official videos on YouTube. This comparison aims to shed light on the importance of having an official video for a song's online presence. The key variable of interest is official\_video, which indicates whether a video is official (1) or non-official (0).

**Chi-Square, ANOVA and Logistic Regression**

The following tests will be inclusive in our analysis for this comparison of YouTube videos:

***Chi-Square Test***: This test was conducted to determine if there is a significant association between the categorical variable official\_video and the engagement metric Views.

***Logistic Regression***: A logistic regression model was built to predict the likelihood of engagement (in terms of comments) based on the official status of the video.

***Data Visualization***: Various plots, including boxplots and bar plots, were created to visually compare the distribution of engagement metrics between official and non-official videos.

**Results**

***Pearson's Chi-squared Test*:**

* + Result: The test statistic was 4042.5 with 3924 degrees of freedom, resulting in a p-value of 0.09145. Since the p-value is greater than the significance level of 0.05, we fail to reject the null hypothesis. There is insufficient evidence to conclude that there is a significant association between the official\_video status and the Likes received.

***Welch Two Sample t-test*:**

* + Result: The test statistic was -13.22 with 3707.4 degrees of freedom, resulting in a p-value less than 2.2e-16. This indicates a significant difference in the mean Views between official and non-official videos.

***ANOVA*:**

* + Result: The F-value was 69.48 with a p-value less than 2e-16, indicating a significant difference in mean Views between the two groups.

***Logistic Regression:***

* + Result: The coefficient estimate for the official\_video variable was statistically significant (p < 0.05), indicating that the official\_video status is a significant predictor of the number of Likes received.

These results suggest that while there is no significant association between the official\_video status and the number of Likes received, there are significant differences in the mean Views between official and non-official videos. Additionally, the official\_video status is a significant predictor of the number of Likes received. These analyses provide a comprehensive understanding of the relationship between video status (official/non-official) and engagement metrics, as well as the predictors that influence the likelihood of a video being official based on its musical features. **A graph with blue squares and green text

Description automatically generated**

*Figure 5A: Mean Engagement Metrics for Official and Non-Official Videos*

**A graph with a blue line

Description automatically generated**

*Figure 5B: Distribution of Views for Official and Non-Official Videos*

**Conclusion/Interpretations**

Based on the analysis, we can conclude that both original t-test and the bootstrapped findings for the album type impact on Spotify streams, suggest that there is enough evidence to reject the null hypothesis but there is not enough evidence to reject the null hypothesis for that of YouTube views in both findings. This means there is likely no difference in average YouTube views between songs released as singles and songs released within albums.

The complex nature of digital music engagement and the limitations of using a single set of features to predict success across platforms. For future analysis, employing machine learning models to incorporate a broader range of variables might offer more insights. Additionally, exploring temporal trends and artist-specific effects could further refine our understanding of what drives engagement on YouTube.

For the impact of tempo and valence on YouTube Likes and Comments, while the analysis provides valuable insights into the relationships between musical attributes and audience engagement metrics, the diagnostic plots suggest potential issues with linearity and normality in the regression models. Further investigation and refinement of the models are warranted to ensure the accuracy and reliability of the analysis results. Additionally, the scatterplots indicate the presence of diverse factors influencing audience engagement metrics, highlighting the complexity of music consumption behavior on platforms like YouTube.

The analysis on the intricate interplay between musical genres and audience engagement on YouTube. Leveraging advanced computational tools and statistical analytics, we uncover key insights into music consumption trends. Through machine learning algorithms, music is effectively categorized into genres, enabling a nuanced analysis of engagement metrics including views, likes, and comments. Our statistical tests highlight significant variations in engagement levels across genres, while regression analysis identifies pivotal factors driving high audience engagement. Additionally, clustering algorithms unveil underlying patterns among genres, shedding light on audience dynamics. These findings offer valuable insights for content creators and marketers, facilitating the optimization of strategies to enhance audience engagement on digital platforms such as YouTube.

Lastly, the analyses collectively suggest that official videos on YouTube tend to attract higher engagement, measured by views, likes, and comments, compared to non-official videos. Additionally, audio features such as danceability, energy, speechiness, and valence play a significant role in determining the official status of a video. These insights can inform content creators and marketers about the importance of producing official videos to enhance their online presence and engagement metrics.

**References**

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

Bluman, A. (n.d.). Elementary Statistics 10th Edition. McGraw Hill ISBN 978-130-7494-327

Kabacoff, R. (n.d.). R in Action 2nd Edition. Manning Publisher ISBN 978-161-7291-388

Tyre, D. (2017, Aug 11). Checking Assumptions of the Linear Model. <https://www.youtube.com/watch?v=E27HcS9QaT0>

The Organic Chemistry Tutor. (2019, Jan 4). Descriptive Statistics vs Inferential Statistics. <https://www.youtube.com/watch?v=VHYOuWu9jQI>

Mack, C. (2016, Oct 24). Lecture46 (Data2Decision) Best Subset Regression in R. <https://www.youtube.com/watch?v=oevjpFdZqmk>

R Programming 101. (2022, Apr 13). Chi Squared Test using R programming. <https://www.youtube.com/watch?v=KubK6hgMbvg>

R Programming 101. (2022, Oct 5). ANOVA using R programming. <https://www.youtube.com/watch?v=0EA5aPQpqYY>

Equitable Equations. (2023, May 10). Logistic Regression in R. <https://www.youtube.com/watch?v=E7J3M1oYVlc>

**Appendix**

The appendix section lists all R codes used for the analyses in this project.

* data <- read\_csv("Spotify\_Youtube.csv")
* summary(data)
* Mode <- function(x) {
  + ux <- unique(x)
  + ux[which.max(tabulate(match(x, ux)))]
  + }
* replace\_na\_with\_mode <- function(x) {
  + if (is.numeric(x)) {
  + x[is.na(x)] <- mean(x, na.rm = TRUE)
  + } else {
  + x[is.na(x)] <- Mode(x)
  + }
  + return(x)
  + }
  + spotify\_youtube <- data %>%
  + select(-"...1") %>% # Remove the variable "...1"
  + mutate(across(everything(), replace\_na\_with\_mode))
  + summary(spotify\_youtube)
* cor(spotify\_youtube[c('Stream', 'Views')], method = 'pearson')
* subset\_data <- sample\_n(spotify\_youtube, 3500)
* singles\_data <- subset\_data %>%
  + filter(Album\_type == "single")
  + albums\_data <- subset\_data %>%
  + filter(Album\_type == "album")
* t\_test\_streams <- t.test(singles\_data$Stream, albums\_data$Stream)
* t\_test\_views <- t.test(singles\_data$Views, albums\_data$Views)
* cat("T-test for Spotify Streams:\n")
  + print(t\_test\_streams)
* cat("\nT-test for YouTube Views:\n")
  + print(t\_test\_views)
* bar\_data <- data.frame(
  + Metric = rep(c("Spotify Streams", "YouTube Views"), each = 2),
  + Album\_type = rep(c("Single", "Album"), 2),
  + Mean\_Values = c(mean(singles\_data$Stream), mean(albums\_data$Stream),
    - mean(singles\_data$Views), mean(albums\_data$Views))
* ggplot(bar\_data, aes(x = Album\_type, y = Mean\_Values, fill = Metric)) +
  + geom\_bar(stat = "identity", position = "dodge") +
  + labs(title = "Mean Spotify Streams and YouTube Views for Singles and Albums",
    - x = "Album Type",
    - y = "Mean Values",
    - fill = "Metric") +
  + theme\_minimal()
* subset\_data <- subset\_data %>% filter(Album\_type %in% c("single", "album"))
* set.seed(123)
* t\_test\_boot <- function(data, indices) {
  + subset\_data <- data[indices, ]
  + t\_test\_result\_stream <- t.test(subset\_data$Stream ~ subset\_data$Album\_type)
  + t\_test\_result\_view <- t.test(subset\_data$Views ~ subset\_data$Album\_type)
  + return(c(
  + stream\_statistic = t\_test\_result\_stream$statistic,
  + stream\_p\_value = t\_test\_result\_stream$p.value,
  + view\_statistic = t\_test\_result\_view$statistic,
  + view\_p\_value = t\_test\_result\_view$p.value
  + ))
  + }
* boot\_results <- boot(data = subset\_data, statistic = t\_test\_boot, R = 1000)
* cat("Bootstrapped Results for Spotify Streams and YouTube Views (Subset):\n")
* print(boot\_results)
* # Read the dataset
* ```{r}
* library(readr)
* Spotify\_Youtube <- read\_csv("~/Winter 2024 1st Quarter/Module 4/Final Project Assignment/Spotify\_Youtube.csv")
* View(Spotify\_Youtube)
* ```
* # 1. Remove the irrelevant column
* ```{r}
* data <- select(Spotify\_Youtube, -1)
* ```
* # 2. Impute missing values
* # For numeric variables, replace NAs with mean
* ```{r}
* numeric\_cols <- sapply(data, is.numeric)
* data[numeric\_cols] <- lapply(data[numeric\_cols], function(x) ifelse(is.na(x), mean(x, na.rm = TRUE), x))
* ```
* # For categorical variables, replace NAs with mode (most common value)
* ```{r}
* getmode <- function(v) {
* uniqv <- unique(v)
* uniqv[which.max(tabulate(match(v, uniqv)))]
* }
* categorical\_cols <- sapply(data, is.factor) | sapply(data, is.character)
* data[categorical\_cols] <- lapply(data[categorical\_cols], function(x) ifelse(is.na(x), getmode(x), x))
* ```
* # 3. Use a subset of rows
* # Assuming you want to work with a specific number of rows, e.g., the first 1000 rows
* ```{r}
* subset\_data <- data[1:1000, ]
* ```
* # Display the first few rows of the processed dataset
* ```{r}
* head(subset\_data)
* ```
* # 4.Select the relevant columns for Spotify metrics and YouTube engagement metrics
* ```{r}
* spotify\_youtube\_metrics <- subset\_data[, c("Danceability", "Energy", "Views", "Likes")]
* ```
* # 5. Compute the Pearson correlation matrix
* ```{r}
* correlation\_matrix <- cor(spotify\_youtube\_metrics, use = "complete.obs") # 'complete.obs' handles missing values by using complete observations
* ```
* ```{r}
* correlation\_matrix
* ```
* # 6. Create scatterplots
* ```{r}
* library(ggplot2)
* ```
* # Scatter plot for Danceability vs. Views
* ```{r}
* ggplot(data, aes(x = Danceability, y = Views)) +
* geom\_point(alpha = 0.5) + # Add points with some transparency
* geom\_smooth(method = "lm", color = "blue", se = FALSE) + # Add a linear regression line without confidence interval
* labs(title = "Danceability vs. YouTube Views",
* x = "Danceability",
* y = "YouTube Views") +
* theme\_minimal()
* ```
* # Scatter plot for Energy vs. Views
* ```{r}
* ggplot(data, aes(x = Energy, y = Views)) +
* geom\_point(alpha = 0.5) + # Add points with some transparency
* geom\_smooth(method = "lm", color = "red", se = FALSE) + # Add a linear regression line without confidence interval
* labs(title = "Energy vs. YouTube Views",
* x = "Energy",
* y = "YouTube Views") +
* theme\_minimal()
* ```
* # View the structure of the dataset
* str(data)
* # Summary statistics of the dataset
* summary(data)
* # Scatter plot of Tempo vs. Likes
* ggplot(data, aes(x = Tempo, y = Likes)) +
* geom\_point() +
* ggtitle("Scatter plot of Tempo vs. Likes") +
* xlab("Tempo") +
* ylab("Likes")
* # Scatter plot of Valence vs. Likes
* ggplot(data, aes(x = Valence, y = Likes)) +
* geom\_point() +
* ggtitle("Scatter plot of Valence vs. Likes") +
* xlab("Valence") +
* ylab("Likes")
* # Scatter plot of Tempo vs. Comments
* ggplot(data, aes(x = Tempo, y = Comments)) +
* geom\_point() +
* ggtitle("Scatter plot of Tempo vs. Comments") +
* xlab("Tempo") +
* ylab("Comments")
* # Scatter plot of Valence vs. Comments
* ggplot(data, aes(x = Valence, y = Comments)) +
* geom\_point() +
* ggtitle("Scatter plot of Valence vs. Comments") +
* xlab("Valence") +
* ylab("Comments")
* # Simple linear regression model for Likes
* likes\_model <- lm(Likes ~ Tempo + Valence, data = data)
* summary(likes\_model)
* # Simple linear regression model for Comments
* comments\_model <- lm(Comments ~ Tempo + Valence, data = data)
* summary(comments\_model)
* # Comparison plot for Likes and Comments regression coefficients
* par(mfrow=c(1,2))
* plot(likes\_model)
* plot(comments\_model)
* data <- data.frame(
* Genre = c("Feel Good Inc.", "Rhinestone Eyes", "New Gold", "On Melancholy Hill", "Clint Eastwood", "DARE", "New Gold (Remix)", "She's My Collar", "Cracker Island"),
* Views = c(693555221, 72011645, 8435055, 211754952, 618480958, 259021161, 451996, 1010982, 24459820),
* Likes = c(6220896, 1079128, 282142, 1788577, 6197318, 1844658, 11686, 17675, 739527),
* Comments = c(169907, 31003, 7399, 55229, 155930, 72008, 241, 260, 20296)
* )
* data\_engagement <- data %>%
* group\_by(genre) %>%
* summarise(total\_engagement = sum(views + likes + comments))
* # Create the bar plot using ggplot2
* ggplot(data\_engagement, aes(x = genre, y = total\_engagement)) +
* geom\_bar(stat = "identity", fill = "skyblue") +
* labs(title = "Total Engagement by Genre", x = "Genre", y = "Total Engagement (Views + Likes + Comments)") +
* theme\_bw()
* # Calculate total engagement (sum of views, likes, and comments)
* data$total\_engagement <- data$Views + data$Likes + data$Comments
* # Calculate total engagement (sum of views, likes, and comments)
* data$total\_engagement <- data$Views + data$Likes + data$Comments
* # Bar plot for total engagement
* barplot\_data <- data.frame(Genre = data$Genre, Total\_Engagement = data$total\_engagement)
* barplot <- ggplot(barplot\_data, aes(x = Genre, y = Total\_Engagement)) +
* geom\_bar(stat = "identity", fill = "skyblue") +
* theme(axis.text.x = element\_text(angle = 45, hjust = 1)) +
* labs(title = "Total Engagement by Genre", x = "Genre", y = "Total Engagement") +
* theme(plot.title = element\_text(hjust = 0.5))
* # Pie chart for engagement breakdown
* pie\_data <- data.frame(
* Engagement = c("Views", "Likes", "Comments"),
* Total = c(sum(data$Views), sum(data$Likes), sum(data$Comments))
* )
* pie\_chart <- ggplot(pie\_data, aes(x = "", y = Total, fill = Engagement)) +
* geom\_bar(stat = "identity", width = 1) +
* coord\_polar("y", start = 0) +
* labs(title = "Engagement Breakdown", fill = "Engagement Type") +
* theme\_void() +
* theme(legend.title = element\_blank())
* # Display plots
* print(barplot)
* print(pie\_chart)
* # Calculate total engagement for each type (views, likes, comments)
* total\_engagement\_by\_type <- data %>%
* summarise(
* views = sum(views),
* likes = sum(likes),
* comments = sum(comments)
* ) %>%
* mutate(total\_engagement = sum(c(views, likes, comments)))
* # Create the pie chart using ggplot2
* ggplot(total\_engagement\_by\_type, aes(x = "", fill = engagement\_type, label = paste0(engagement\_type, ": ", .value))) +
* geom\_pie(radius = 1.5, angle = 90) +
* coord\_polar(theta = "circular") +
* labs(title = "Distribution of Engagement Types", fill = "Engagement Type") +
* theme\_bw()
* str(spotify\_youtube)
* # Clean and preprocess the data
* # Select relevant columns and mutate 'official\_video' column
* spotify\_youtube\_clean <- spotify\_youtube %>%
* mutate(official\_video = ifelse(official\_video == "TRUE", 1, 0)) %>%
* dplyr::select(Track, Artist, Views, Likes, Comments, official\_video)
* # Descriptive statistics
* summary(spotify\_youtube\_clean)
* #Impute missing values with the median
* spotify\_youtube\_clean <- spotify\_youtube\_clean %>%
* mutate\_if(is.numeric, ~ ifelse(is.na(.), median(., na.rm = TRUE), .))
* # Set seed for reproducibility
* set.seed(123)
* # Sample 20% of the data
* sampled\_data <- spotify\_youtube\_clean[sample(nrow(spotify\_youtube\_clean), 0.2 \* nrow(spotify\_youtube\_clean)), ]
* # Load the subset of the dataset
* spotify\_youtube\_subset <- sampled\_data
* # Check the structure of the dataset
* str(spotify\_youtube\_subset)
* # Calculate mean views, likes, and comments for official and non-official videos
* mean\_engagement <- spotify\_youtube\_subset %>%
* group\_by(official\_video) %>%
* summarise(mean\_views = mean(Views),
* mean\_likes = mean(Likes),
* mean\_comments = mean(Comments))
* # Convert data to long format for plotting
* mean\_engagement\_long <- pivot\_longer(mean\_engagement, cols = c(mean\_views, mean\_likes, mean\_comments),
* names\_to = "Engagement Metric", values\_to = "Mean Value")
* # Create bar plot
* ggplot(mean\_engagement\_long, aes(x = official\_video, y = `Mean Value`, fill = `Engagement Metric`)) +
* geom\_bar(stat = "identity", position = "dodge") +
* labs(x = "Official Video", y = "Mean Value", title = "Mean Engagement Metrics for Official and Non-Official Videos") +
* theme\_minimal() +
* theme(legend.position = "top")
* #Box Plot for Views, Likes, and Comments:
* # Create box plot
* ggplot(spotify\_youtube\_subset, aes(x = factor(official\_video), y = Views, fill = factor(official\_video))) +
* geom\_boxplot() +
* labs(x = "Official Video", y = "Views", title = "Distribution of Views for Official and Non-Official Videos") +
* scale\_fill\_manual(values = c("orange", "blue"), labels = c("Non-Official", "Official")) +  # Specify labels for the legend
* theme\_minimal()
* # Perform statistical tests (Chi-Square, ANOVA, GLM, Logistic Regression) as per the research question
* # Chi-Square test for engagement metrics between official and non-official videos
* chi\_square\_result <- chisq.test(spotify\_youtube\_subset$official\_video, spotify\_youtube\_subset$Likes)
* print(chi\_square\_result)
* # Perform a two-sample t-test
* t\_test\_result <- t.test(Views ~ official\_video, data = spotify\_youtube\_subset)
* # Print the results
* print(t\_test\_result)
* # ANOVA for engagement metrics between official and non-official videos
* anova\_result\_views <- aov(Views ~ official\_video, data = spotify\_youtube\_subset)
* print(summary(anova\_result\_views))
* # GLM for engagement metrics between official and non-official videos
* glm\_model\_likes <- glm(Likes ~ official\_video, data = spotify\_youtube\_subset, family = "gaussian")
* print(summary(glm\_model\_likes))
* # Logistic Regression for predicting official\_video
* logistic\_model <- glm(official\_video ~ Danceability + Energy + Speechiness + Acousticness + Liveness + Valence + Tempo,
* data = spotify\_youtube, family = binomial)
* # Summary of the logistic regression model
* summary(logistic\_model)