Group Project part 1

Aleksandra Tatko (8022399) Clarine Schornagel (7551010)
Nikita Smirnov (0399515) Filippo Sallustio (8858926)

2024-06-05

Contents

1. Introduction

The aim of this project is to analyze a dataset of popular songs from Spotify to investigate the hypothesis that the music vibe, particularly the danceability of songs, has remained consistent over the years. The dataset, sourced from Kaggle and titled "Spotify - Top 2000", includes a range of features such as year of release, beats per minute (BPM), danceability, and popularity, among others.

For our research, we focused on the variables year of release, BPM, danceability, and popularity. We began our analysis by cleaning the dataset and extracting the relevant variables to streamline the subsequent analysis. To facilitate our research, we grouped the years into decades and identified the most popular genre for each decade. This categorization helped us establish a foundation for analyzing trends in danceability over time and to test our hypothesis comprehensively.

```
# read the file and then show top 10 rows just to check if everything works fine
library(dplyr)
library(readr)

spoty <- read_csv("Spotify-2000.csv")
head(spoty, 10)
```

```
# A tibble: 10 x 15
   Index Title           Artist `Top Genre`  Year Beats Per Minute (BP~1 Energy
   <dbl> <chr>           <chr>  <chr>       <dbl>                  <dbl>  <dbl>
 1     1 Sunrise         Norah~ adult stan~  2004                    157     30
 2     2 Black Night     Deep ~ album rock   2000                    135     79
```

```
 3     3 Clint Eastwood   Goril~ alternativ~  2001                   168     69
 4     4 The Pretender    Foo F~ alternativ~  2007                   173     96
 5     5 Waitin' On A Su~ Bruce~ classic ro~  2002                   106     82
 6     6 The Road Ahead ~ City ~ alternativ~  2004                    99     46
 7     7 She Will Be Lov~ Maroo~ pop          2002                   102     71
 8     8 Knights of Cydo~ Muse   modern rock  2006                   137     96
 9     9 Mr. Brightside   The K~ modern rock  2004                   148     92
10    10 Without Me       Eminem detroit hi~  2002                   112     67
# i abbreviated name: 1: `Beats Per Minute (BPM)`
# i 8 more variables: Danceability <dbl>, `Loudness (dB)` <dbl>,
#   Liveness <dbl>, Valence <dbl>, `Length (Duration)` <dbl>,
#   Acousticness <dbl>, Speechiness <dbl>, Popularity <dbl>
```

```r
# Creating a new column 'Decade' to categorize each year into decades
cleaned_spoty <- cleaned_spoty %>%
  mutate(Decade = floor(Year / 10) * 10)

# Find the most popular genre for each decade
most_popular_genre <- cleaned_spoty %>%
  group_by(Decade, Genre) %>%
  summarise(Average_Popularity = mean(Popularity, na.rm = TRUE)) %>%
  arrange(Decade, desc(Average_Popularity)) %>%
  slice(1) %>%
  ungroup()

# Rename the columns for clarity
most_popular_genre <- most_popular_genre %>%
  rename(Most_Popular_Genre = Genre, Most_Popular_Genre_Avg_Popularity = Average_Popularity)

# Join the most popular genre back to the original dataset
cleaned_spoty <- cleaned_spoty %>%
  left_join(most_popular_genre, by = "Decade")

# View the updated dataset
head(cleaned_spoty)
```

```
# A tibble: 6 x 8
  Genre               Year Beats Per Minute (BP~1 Danceability Popularity Decade
  <chr>              <dbl>                  <dbl>        <dbl>      <dbl>  <dbl>
1 adult standards     2004                    157           53         71   2000
2 album rock          2000                    135           50         39   2000
3 alternative hip h~  2001                    168           66         69   2000
4 alternative metal   2007                    173           43         76   2000
5 classic rock        2002                    106           58         59   2000
6 alternative pop r~  2004                     99           54         45   2000
# i abbreviated name: 1: `Beats Per Minute (BPM)`
# i 2 more variables: Most_Popular_Genre <chr>,
#   Most_Popular_Genre_Avg_Popularity <dbl>
```

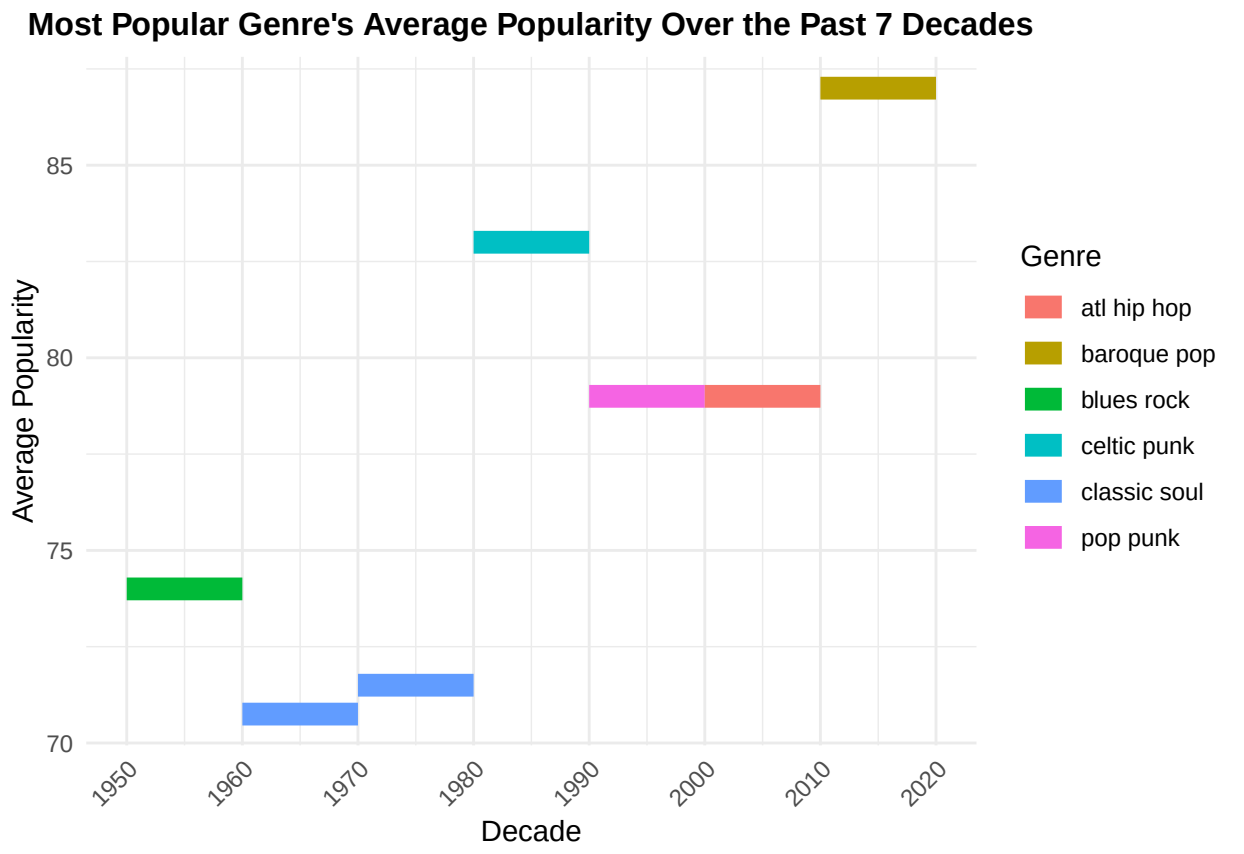
---

2. Visualization

To understand the structure of our data and identify any potential trends, we started with the following visualization that shows genres and their popularity in each decade. This visualization serves as a foundational step in testing our hypothesis that "The music vibe hasn't changed over the years," specifically focusing on danceability. By first identifying the most popular genres in each decade and their popularity, we set the stage for a deeper analysis into whether the danceability of songs within these genres has remained consistent over time.

```
# graphs start here
library(dplyr)
library(ggplot2)

ggplot(most_popular_genre, aes(x = Decade, y = Most_Popular_Genre_Avg_Popularity)) +
  geom_segment(aes(x = Decade, xend = Decade + 10, y = Most_Popular_Genre_Avg_Popularity, yend = Most_P
  labs(title = "Most Popular Genre's Average Popularity Over the Past 7 Decades",
       x = "Decade",
       y = "Average Popularity",
       color = "Genre") +
  scale_x_continuous(breaks = seq(min(most_popular_genre$Decade), max(most_popular_genre$Decade) + 10,
                     limits = c(min(most_popular_genre$Decade), 2020)) +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1),
        plot.title = element_text(size = 12, face = "bold", hjust = 0.5))
```



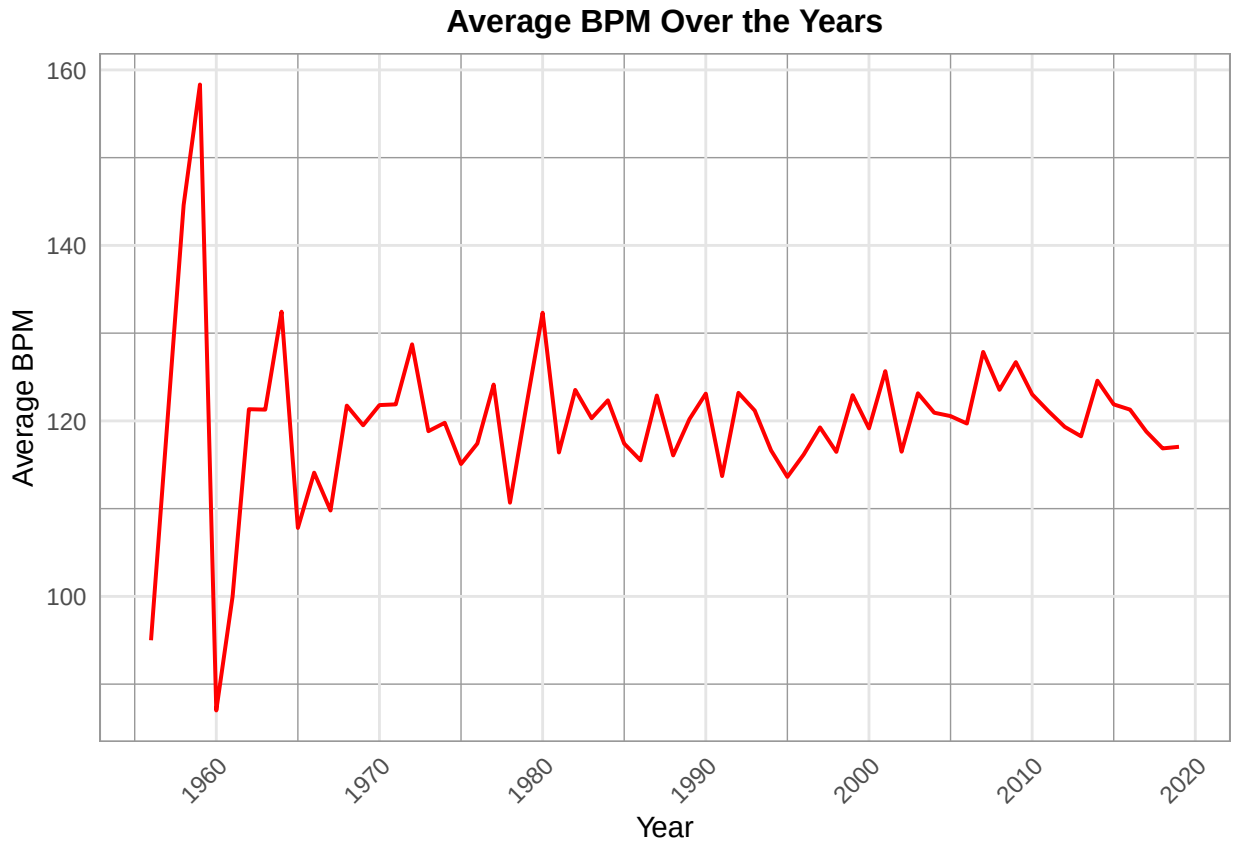
The graph indicates that while the most popular genre varies across decades, the average popularity scores

remain relatively stable, providing a preliminary insight that the music vibe, as measured by popularity, has not drastically shifted. The next steps will involve a detailed analysis of the danceability scores within these genres across the decades to further substantiate our hypothesis.

As our next step, we decided to examine the average BPM and danceability of popular songs over several decades to further investigate our hypothesis. The graphs of Average BPM and Average Danceability over the years provide valuable insights. Both graphs reveal significant fluctuations in the 1960s, with the average BPM and danceability scores showing considerable variability. However, from the 1970s onwards, these trends stabilize.

```
average_bpm_per_year <- cleaned_spoty %>%
  group_by(Year) %>%
  summarise(Average_BPM = mean(`Beats Per Minute (BPM)`, na.rm = TRUE))

ggplot(average_bpm_per_year, aes(x = Year, y = Average_BPM)) +
  geom_line(color = "red", size = 0.7) +
  labs(title = "Average BPM Over the Years",
       x = "Year",
       y = "Average BPM") +
  scale_x_continuous(breaks = seq(1950, 2020, by = 10)) +
  theme_minimal() +
  theme(
    axis.text.x = element_text(angle = 45, hjust = 1),
    plot.title = element_text(size = 12, face = "bold", hjust = 0.5),
    panel.grid.major = element_line(size = 0.5, linetype = 'solid', colour = "gray90"),
    panel.grid.minor = element_line(size = 0.25, linetype = 'solid', colour = "gray60"),
    panel.border = element_rect(color = "gray60", fill = NA, size = 0.5)
  )
```

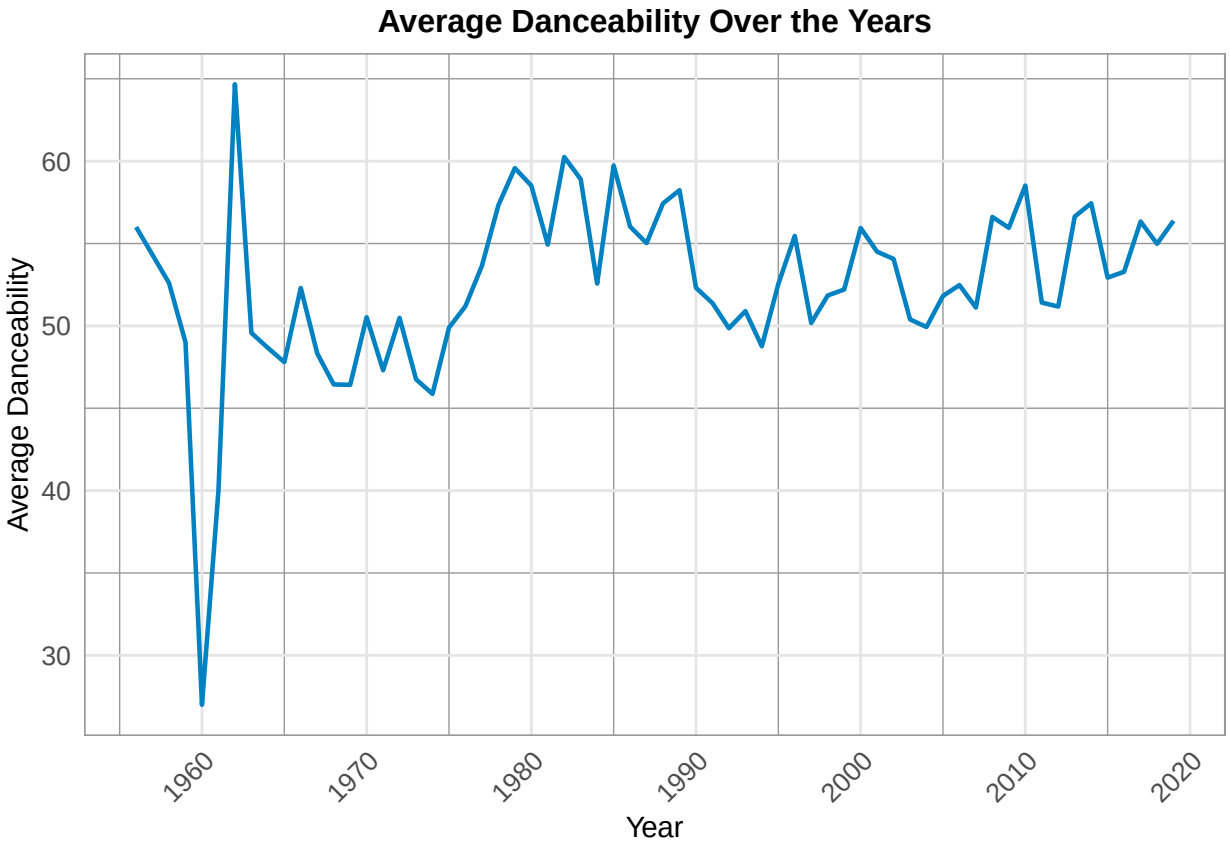


The Average BPM graph indicates that from the 1970s, BPM values maintain a relatively consistent range, fluctuating around 120 BPM with only minor variations. This suggests that the tempo of popular music has remained fairly stable over the decades, with no clear upward or downward trend.

```
library(dplyr)
library(ggplot2)

average_danceability_per_year <- cleaned_spoty %>%
  group_by(Year) %>%
  summarise(Average_Danceability = mean(Danceability, na.rm = TRUE))

ggplot(average_danceability_per_year, aes(x = Year, y = Average_Danceability)) +
  geom_line(color = "#0082C2", size = 0.8) +
  labs(title = "Average Danceability Over the Years",
       x = "Year",
       y = "Average Danceability") +
  scale_x_continuous(breaks = seq(1950, 2020, by = 10)) +
  theme_minimal() +
  theme(
    axis.text.x = element_text(angle = 45, hjust = 1, size = 10),
    axis.text.y = element_text(size = 10),
    plot.title = element_text(size = 12, face = "bold", hjust = 0.5),
    panel.grid.major = element_line(size = 0.5, linetype = 'solid', colour = "gray90"),
    panel.grid.minor = element_line(size = 0.25, linetype = 'solid', colour = "gray60"),
    panel.border = element_rect(color = "gray60", fill = NA, size = 0.5))
```



On the other hand, the Average Danceability graph shows a general upward trend from the 1970s onwards, stabilizing around a danceability score of 60. While there are some minor peaks and troughs, the overall trend indicates a slight increase in danceability over time. However, these changes are not drastic, further supporting the notion that the core aspects of the music vibe, particularly tempo and danceability, have remained relatively consistent over the years.

---

3. Linear Regression

As the last step in our analysis, we sought to determine the significance of our findings by performing linear regression to assess whether the Decade variable significantly affects Danceability. If the coefficients for the Decade variable are not significantly different from zero, it would suggest that the music vibe, as measured by Danceability, has not changed significantly over the years.

We split the data into training and test sets and performed both Lasso and Ridge regression. The Mean Squared Error (MSE) for both models was relatively similar, with the Ridge Regression MSE being 221.85 and the Lasso Regression MSE being 221.51. This similarity in MSE indicates that the models are performing comparably, and there is no significant difference in predictive performance between the two methods.

Both Ridge and Lasso regression analyses yielded coefficients for the Decade variable that were not significantly different from zero. This outcome suggests that the music vibe, as measured by Danceability, has remained relatively stable over the years. Consequently, our hypothesis that the music vibe has not changed significantly over time is supported by the regression analysis, as there are no substantial shifts in Danceability attributable to different decades.

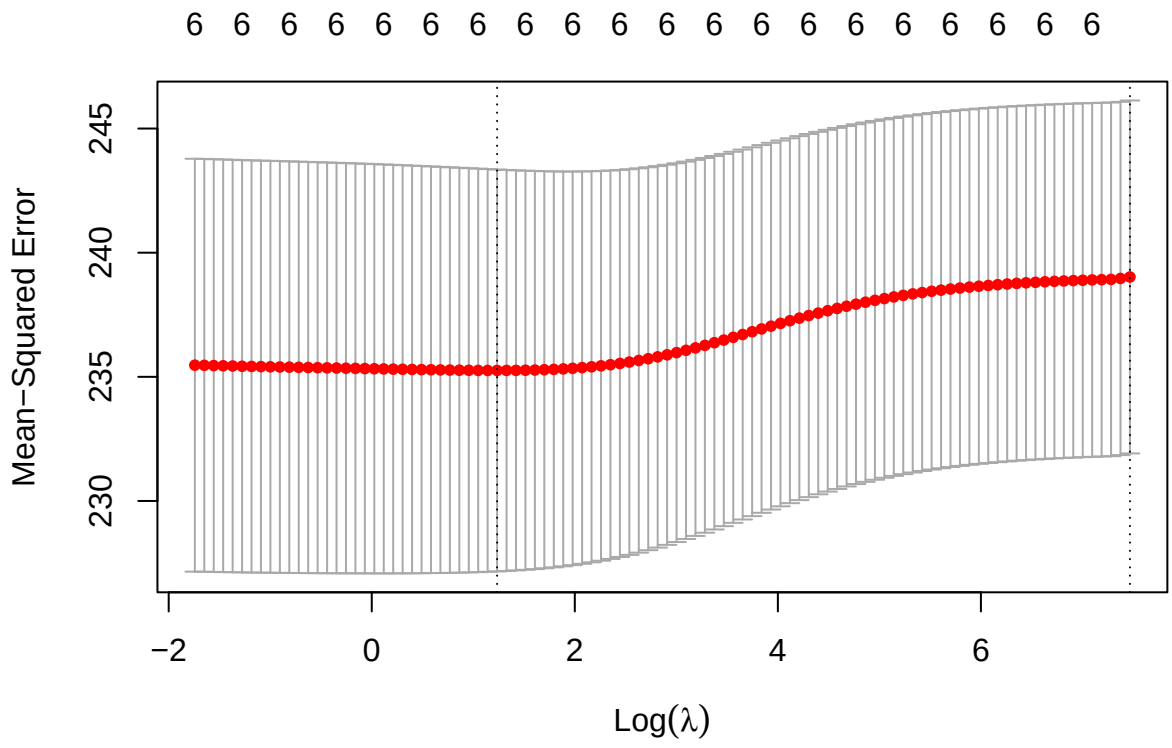
```r
library(glmnet)
library(gridExtra)
library(caret)

# Prepare the data for regression
cleaned_spoty$Decade <- as.factor(cleaned_spoty$Decade)

# Create model matrix for regression
X <- model.matrix(Danceability ~ Decade, data = cleaned_spoty)[, -1]
y <- cleaned_spoty$Danceability

# Split data into training and test sets
set.seed(123)
trainIndex <- createDataPartition(y, p = 0.7, list = FALSE)
X_train <- X[trainIndex, ]
X_test <- X[-trainIndex, ]
y_train <- y[trainIndex]
y_test <- y[-trainIndex]

# Ridge Regression
ridge_model <- cv.glmnet(X_train, y_train, alpha = 0)
ridge_best_lambda <- ridge_model$lambda.min
plot(ridge_model)
```



```r
# Predict and evaluate
ridge_predictions <- predict(ridge_model, s = ridge_best_lambda, newx = X_test)
ridge_mse <- mean((ridge_predictions - y_test)^2)
print(paste("Ridge Regression MSE:", ridge_mse))
```

```
[1] "Ridge Regression MSE: 221.854767390612"
```

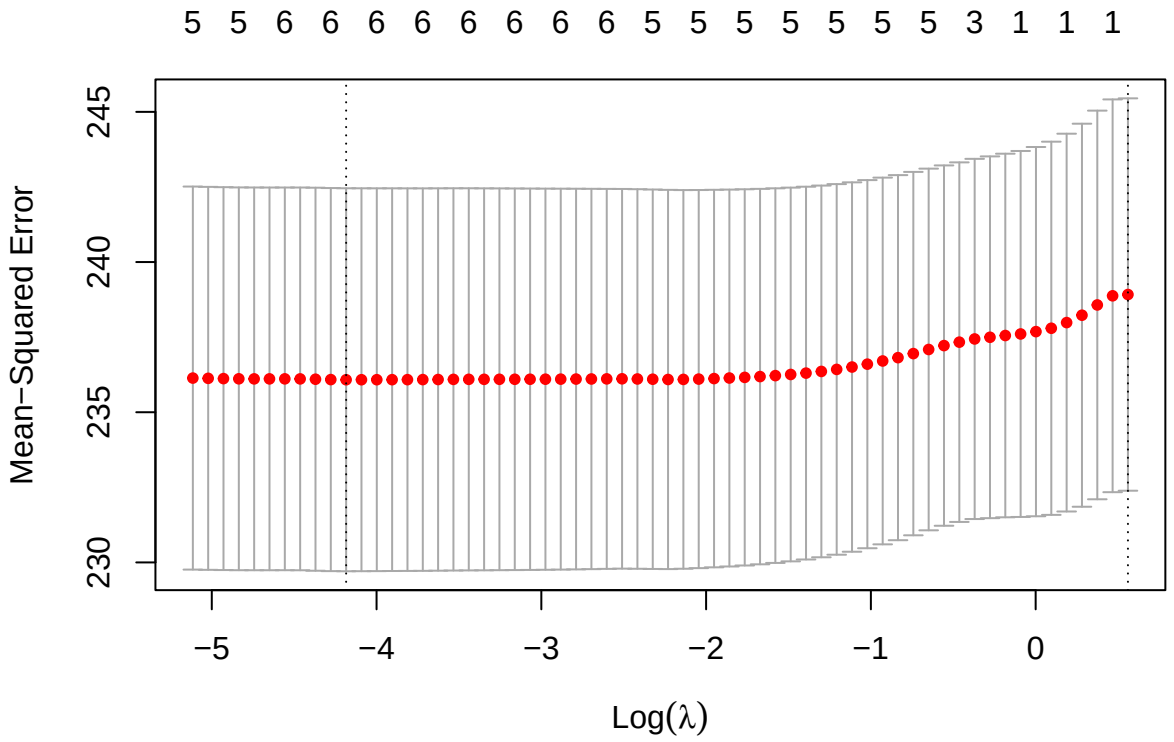
```r
# Coefficients
ridge_coef <- coef(ridge_model, s = ridge_best_lambda)
print("Ridge Regression Coefficients:")
```

```
[1] "Ridge Regression Coefficients:"
```

```r
print(ridge_coef)
```

```
7 x 1 sparse Matrix of class "dgCMatrix"
                    s1
(Intercept) 53.2158309
Decade1960  -2.6550106
Decade1970  -1.8587155
Decade1980   3.3316204
Decade1990  -1.3831847
Decade2000   0.4992291
Decade2010   0.9923284
```

```r
# Lasso Regression
lasso_model <- cv.glmnet(X_train, y_train, alpha = 1)
lasso_best_lambda <- lasso_model$lambda.min
plot(lasso_model)
```




```
# Predict and evaluate
lasso_predictions <- predict(lasso_model, s = lasso_best_lambda, newx = X_test)
lasso_mse <- mean((lasso_predictions - y_test)^2)
print(paste("Lasso Regression MSE:", lasso_mse))
```

```
[1] "Lasso Regression MSE: 221.50504877157"
```

```
# Coefficients
lasso_coef <- coef(lasso_model, s = lasso_best_lambda)
print("Lasso Regression Coefficients:")
```

```
[1] "Lasso Regression Coefficients:"
```

```
print(lasso_coef)
```

```
7 x 1 sparse Matrix of class "dgCMatrix"
                    s1
(Intercept) 51.8040208
Decade1960  -1.7422481
Decade1970  -0.7569830
Decade1980   5.3207110
Decade1990  -0.1970654
Decade2000   1.9675517
Decade2010   2.5506727
```

4. Conclusion

In conclusion, our analysis supports the hypothesis that the music vibe, specifically the danceability of songs, has remained consistent over the decades. Through the use of visualizations and regression analysis, we observed that while there are some fluctuations in BPM and danceability scores, especially in the 1960s, these trends stabilize from the 1970s onwards. The regression analysis further confirmed that the Decade variable does not significantly affect Danceability, indicating no substantial change in the music vibe over time.

These findings suggest that the core aspects of music that contribute to its danceability have remained stable, despite changes in genre popularity and other musical elements. This stability can be valuable information for music producers and artists aiming to understand long-term trends in music preferences. Future research could expand on this analysis by incorporating additional variables and exploring non-linear relationships to gain a more comprehensive understanding of trends in music characteristics over time.

5. Appendix

The responsibilities of each group member: Nikita Smirnov: cleaning the data, choosing the research question, preparing the files; Clarine Schornagel: Creating the graphs for data visulization; Filippo Sallustio: Creating the linear regression graphs; Aleksandra Tatko: Writing the report.