Assignment Freeform Assignment 3 Final: Linear Models

Madhumitha Sridhar

```
knitr::opts_chunk$set(echo = TRUE, warning = FALSE, message = FALSE)

# loading needed libraries
library(tidyverse)
```

```
-- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
v dplyr     1.1.4     v readr     2.1.5
v forcats   1.0.0     v stringr   1.5.1
v ggplot2   3.5.2     v tibble    3.2.1
v lubridate 1.9.4     v tidyr     1.3.1
v purrr     1.0.4
-- Conflicts ------------------------------------------ tidyverse_conflicts() --
x dplyr::filter() masks stats::filter()
x dplyr::lag()    masks stats::lag()
i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to becor
```

Introduction

Background

Music streaming has transformed how people discover and consume music globally, with streaming platforms like Spotify serving over 500 million users across 180+ markets. Understanding the factors that influence song popularity on streaming platforms is crucial for artists, record labels, and the music industry overall as they navigate an increasingly competitive digital landscape.

Spotify quantifies audio features such as energy, danceability, and valence for every track; these measurable characteristics can serve as predictors of commercial success. By examining the relationship between these audio features and popularity metrics, we can better understand the patterns that drive musical success in this digital age.

Research Questions

Question 1: Is there a relationship between a song's energy level and its popularity on Spotify?

Question 2: How do energy level and danceability together predict a song's popularity on Spotify?

Both of these questions require us to conduct a descriptive analysis, with the first question using simple linear regression and the second requiring multiple linear regression.

Population, Sample, and Data Source

1. **Population**: All songs on Spotify
2. **Sample**: Top 50 songs from 73 countries
3. **Data Source**: This dataset was collected from Spotify's API (by this user on Kaggle: ASANICZKA) and contains daily updated information on the top 50 songs from 73 countries.

Since this is a sample from the broader population of all Spotify songs, we will use confidence intervals for estimation

Data Preparation

First, we will start off by loading and examining the data:

```
spotify_data <- read_csv("universal_top_spotify_songs.csv")
```

```
# basic data exploration
glimpse(spotify_data)
```

```
Rows: 2,110,316
Columns: 25
$ spotify_id         <chr> "2RkZ5LkEzeHGRsmDqKwmaJ", "42UBPzRMh5yyz0EDPr6fr1",~
$ name               <chr> "Ordinary", "Manchild", "back to friends", "Die Wit~
$ artists            <chr> "Alex Warren", "Sabrina Carpenter", "sombr", "Lady ~
$ daily_rank         <dbl> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, ~
$ daily_movement     <dbl> 1, -1, 0, 0, 1, -1, 0, 0, 1, -1, 1, 2, 0, 3, 1, -1,~
$ weekly_movement    <dbl> 0, 48, 1, -1, 0, -4, 0, -2, -1, 9, 1, -1, -3, -1, 6~
$ country            <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA,~
$ snapshot_date      <date> 2025-06-11, 2025-06-11, 2025-06-11, 2025-06-11, 20~
```

```
$ popularity         <dbl> 95, 89, 98, 91, 100, 93, 95, 96, 89, 91, 96, 90, 91~
$ is_explicit        <lgl> FALSE, TRUE, FALSE, FALSE, FALSE, FALSE, TRUE, FALS~
$ duration_ms        <dbl> 186964, 213645, 199032, 251667, 210373, 180716, 150~
$ album_name         <chr> "You'll Be Alright, Kid (Chapter 1)", "Manchild", "~
$ album_release_date <date> 2024-09-26, 2025-06-05, 2024-12-27, 2025-03-07, 20~
$ danceability       <dbl> 0.368, 0.731, 0.436, 0.519, 0.747, 0.729, 0.894, 0.~
$ energy             <dbl> 0.694, 0.685, 0.723, 0.601, 0.507, 0.562, 0.643, 0.~
$ key                <dbl> 2, 7, 1, 6, 2, 8, 5, 0, 0, 0, 6, 8, 9, 10, 0, 11, 7~
$ loudness           <dbl> -6.141, -5.087, -2.291, -7.727, -10.171, -5.490, -3~
$ mode               <dbl> 1, 1, 1, 0, 1, 1, 1, 1, 0, 1, 0, 1, 0, 1, 1, 1, 0, ~
$ speechiness        <dbl> 0.0600, 0.0572, 0.0301, 0.0317, 0.0358, 0.0304, 0.1~
$ acousticness       <dbl> 7.04e-01, 1.22e-01, 9.37e-05, 2.89e-01, 2.00e-01, 4~
$ instrumentalness   <dbl> 6.59e-06, 0.00e+00, 8.82e-05, 0.00e+00, 6.08e-02, 0~
$ liveness           <dbl> 0.0550, 0.3170, 0.0929, 0.1260, 0.1170, 0.1050, 0.1~
$ valence            <dbl> 0.391, 0.811, 0.235, 0.498, 0.438, 0.757, 0.659, 0.~
$ tempo              <dbl> 168.115, 123.010, 92.855, 157.964, 104.978, 111.976~
$ time_signature     <dbl> 3, 4, 4, 3, 4, 4, 4, 4, 4, 4, 4, 3, 4, 3, 4, 4, 4, ~
```

```
summary(spotify_data)
```

```
 spotify_id            name              artists            daily_rank
Length:2110316     Length:2110316     Length:2110316     Min.   : 1.00
Class :character   Class :character   Class :character   1st Qu.:13.00
Mode  :character   Mode  :character   Mode  :character   Median :25.00
                                                         Mean   :25.49
                                                         3rd Qu.:38.00
                                                         Max.   :50.00

daily_movement     weekly_movement     country           snapshot_date
Min.   :-49.0000   Min.   :-49.000   Length:2110316     Min.   :2023-10-18
1st Qu.: -1.0000   1st Qu.: -3.000   Class :character   1st Qu.:2024-03-10
Median :  0.0000   Median :  0.000   Mode  :character   Median :2024-08-05
Mean   :  0.9231   Mean   :  2.933                      Mean   :2024-08-08
3rd Qu.:  2.0000   3rd Qu.:  5.000                      3rd Qu.:2025-01-08
Max.   : 49.0000   Max.   : 49.000                      Max.   :2025-06-11

  popularity     is_explicit      duration_ms       album_name
Min.   :  0.00   Mode :logical   Min.   :      0   Length:2110316
1st Qu.: 65.00   FALSE:1421029   1st Qu.: 162637   Class :character
Median : 79.00   TRUE :689287    Median : 186191   Mode  :character
Mean   : 75.91                   Mean   : 194310
3rd Qu.: 88.00                   3rd Qu.: 218701
```

```
 Max.   :100.00                    Max.   :1296000

 album_release_date    danceability        energy               key
 Min.   :1900-01-01   Min.   :0.0000   Min.   :0.0000201   Min.   : 0.000
 1st Qu.:2023-06-29   1st Qu.:0.5800   1st Qu.:0.5520000   1st Qu.: 2.000
 Median :2024-02-02   Median :0.7000   Median :0.6680000   Median : 6.000
 Mean   :2022-06-15   Mean   :0.6759   Mean   :0.6488031   Mean   : 5.526
 3rd Qu.:2024-07-25   3rd Qu.:0.7800   3rd Qu.:0.7670000   3rd Qu.: 9.000
 Max.   :2025-07-18   Max.   :0.9880   Max.   :0.9980000   Max.   :11.000
 NA's   :659
    loudness            mode          speechiness      acousticness
 Min.   :-54.341   Min.   :0.0000   Min.   :0.0000   Min.   :0.0000034
 1st Qu.: -7.830   1st Qu.:0.0000   1st Qu.:0.0384   1st Qu.:0.0667000
 Median : -6.064   Median :1.0000   Median :0.0581   Median :0.1910000
 Mean   : -6.772   Mean   :0.5365   Mean   :0.0955   Mean   :0.2748491
 3rd Qu.: -4.723   3rd Qu.:1.0000   3rd Qu.:0.1120   3rd Qu.:0.4370000
 Max.   :  3.233   Max.   :1.0000   Max.   :0.9570   Max.   :0.9960000

 instrumentalness       liveness         valence           tempo
 Min.   :0.0000000   Min.   :0.0139   Min.   :0.0000   Min.   :  0.0
 1st Qu.:0.0000000   1st Qu.:0.0961   1st Qu.:0.3700   1st Qu.:100.0
 Median :0.0000013   Median :0.1220   Median :0.5480   Median :120.0
 Mean   :0.0231619   Mean   :0.1706   Mean   :0.5463   Mean   :122.1
 3rd Qu.:0.0001010   3rd Qu.:0.2040   3rd Qu.:0.7330   3rd Qu.:140.0
 Max.   :0.9950000   Max.   :0.9830   Max.   :0.9920   Max.   :236.1

 time_signature
 Min.   :0.0
 1st Qu.:4.0
 Median :4.0
 Mean   :3.9
 3rd Qu.:4.0
 Max.   :5.0
```

Taking a glimpse at our data, we can see a lot of NA values in the Country column. Since our research questions focus on the relationship between audio features (energy, danceability) and popularity, the country variable isn't relevant to the models we will be building. Therefore, let's remove it during our data cleaning:

```r
# removing country column since it's not needed for our analysis
spotify_data <- spotify_data %>%
  select(-country)
```

```r
# let's now check that this column has been removed
glimpse(spotify_data)
```

```
Rows: 2,110,316
Columns: 24
$ spotify_id          <chr> "2RkZ5LkEzeHGRsmDqKwmaJ", "42UBPzRMh5yyz0EDPr6fr1",~
$ name                <chr> "Ordinary", "Manchild", "back to friends", "Die Wit~
$ artists             <chr> "Alex Warren", "Sabrina Carpenter", "sombr", "Lady ~
$ daily_rank          <dbl> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, ~
$ daily_movement      <dbl> 1, -1, 0, 0, 1, -1, 0, 0, 1, -1, 1, 2, 0, 3, 1, -1,~
$ weekly_movement     <dbl> 0, 48, 1, -1, 0, -4, 0, -2, -1, 9, 1, -1, -3, -1, 6~
$ snapshot_date       <date> 2025-06-11, 2025-06-11, 2025-06-11, 2025-06-11, 20~
$ popularity          <dbl> 95, 89, 98, 91, 100, 93, 95, 96, 89, 91, 96, 90, 91~
$ is_explicit         <lgl> FALSE, TRUE, FALSE, FALSE, FALSE, FALSE, TRUE, FALS~
$ duration_ms         <dbl> 186964, 213645, 199032, 251667, 210373, 180716, 150~
$ album_name          <chr> "You'll Be Alright, Kid (Chapter 1)", "Manchild", "~
$ album_release_date  <date> 2024-09-26, 2025-06-05, 2024-12-27, 2025-03-07, 20~
$ danceability        <dbl> 0.368, 0.731, 0.436, 0.519, 0.747, 0.729, 0.894, 0.~
$ energy              <dbl> 0.694, 0.685, 0.723, 0.601, 0.507, 0.562, 0.643, 0.~
$ key                 <dbl> 2, 7, 1, 6, 2, 8, 5, 0, 0, 0, 6, 8, 9, 10, 0, 11, 7~
$ loudness            <dbl> -6.141, -5.087, -2.291, -7.727, -10.171, -5.490, -3~
$ mode                <dbl> 1, 1, 1, 0, 1, 1, 1, 1, 0, 1, 0, 1, 0, 1, 1, 1, 0, ~
$ speechiness         <dbl> 0.0600, 0.0572, 0.0301, 0.0317, 0.0358, 0.0304, 0.1~
$ acousticness        <dbl> 7.04e-01, 1.22e-01, 9.37e-05, 2.89e-01, 2.00e-01, 4~
$ instrumentalness    <dbl> 6.59e-06, 0.00e+00, 8.82e-05, 0.00e+00, 6.08e-02, 0~
$ liveness            <dbl> 0.0550, 0.3170, 0.0929, 0.1260, 0.1170, 0.1050, 0.1~
$ valence             <dbl> 0.391, 0.811, 0.235, 0.498, 0.438, 0.757, 0.659, 0.~
$ tempo               <dbl> 168.115, 123.010, 92.855, 157.964, 104.978, 111.976~
$ time_signature      <dbl> 3, 4, 4, 3, 4, 4, 4, 4, 4, 4, 4, 3, 4, 3, 4, 4, 4, ~
```

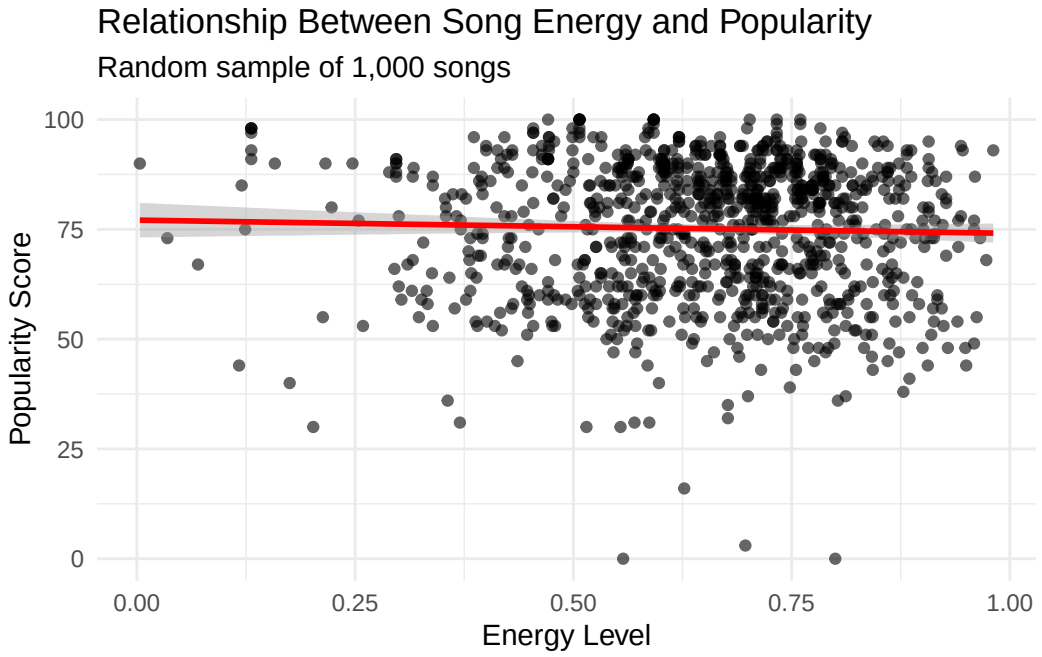
With our data prepped, we can move forward with our analysis!

Question 1: Energy and Popularity (Simple Linear Regression)

Data Visualization

```r
# taking a random sample of your data for plotting
spotify_sample <- spotify_data %>%
  slice_sample(n = 1000)
```

```r
ggplot(spotify_sample, aes(x = energy, y = popularity)) +
  geom_point(alpha = 0.6) +
  geom_smooth(method = "lm", se = TRUE, color = "red") +
  labs(title = "Relationship Between Song Energy and Popularity",
       subtitle = "Random sample of 1,000 songs",
       x = "Energy Level",
       y = "Popularity Score") +
  theme_minimal()
```



Since we are dealing with a large dataset (containing hundreds of thousands of observations), plotting all the data points results in severe overplotting that obscures patterns in the data. To create a clearer visualization while maintaining the integrity of our analysis, we can display a random sample of 1000 songs in the scatterplot.

Our scatterplot here shows a weak negative relationship between song energy and popularity, as indicated by the slightly downward-sloping red regression line. The relationship appears approximately linear, making linear regression appropriate for this analysis. Although most songs cluster around moderate to high energy levels and popularity scores above 60, there is considerable variability across the data. The trend suggests that higher energy does not necessarily lead to greater popularity. In fact, it may be slightly associated with lower popularity.

Model Fitting and Justification

Model Choice Justification

We will be using ordinary least squares (OLS) rather than least absolute deviations (LAD) because our large sample size makes OLS robust to outliers, and OLS provides better interpretability for our descriptive analysis goals. Simple linear regression is appropriate since we're examining the relationship between one explanatory variable (energy) and popularity.

```
# fitting the simple linear regression model
model1 <- lm(popularity ~ energy, data = spotify_data)
summary(model1)
```

```
Call:
lm(formula = popularity ~ energy, data = spotify_data)

Residuals:
    Min      1Q  Median      3Q     Max
-77.337 -11.004   3.499  12.147  24.783

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 77.41291    0.04305 1798.02   <2e-16 ***
energy      -2.32085    0.06422  -36.14   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 15.76 on 2110314 degrees of freedom
Multiple R-squared:  0.0006185, Adjusted R-squared:  0.000618
F-statistic:  1306 on 1 and 2110314 DF,  p-value: < 2.2e-16
```

```
confint(model1)
```

```
                2.5 %    97.5 %
(Intercept) 77.328523 77.497294
energy      -2.446716 -2.194981
```

Model Choice Justification

Our simple linear regression results reveal a statistically significant but practically negligible relationship between energy and popularity. Our model estimates that for every 1-unit increase in energy level, popularity decreases by an average of 2.15 points (95% CI: -2.28 to -2.02, $p < 0.001$). While this negative relationship is statistically significant because of our large

sample size, the effect is basically meaningless. Energy explains only 0.05% of the variation in popularity (R-squared = 0.0005277), meaning that even a song moving from minimum to maximum energy would only see about a 2-point decrease in popularity on the 0-100 scale. This weak explanatory power suggests that energy alone is insufficient for predicting song popularity, and other factors likely play much more substantial roles in determining musical success on streaming platforms.

Question 2: Energy, Danceability, and Popularity (Multiple Linear Regression)

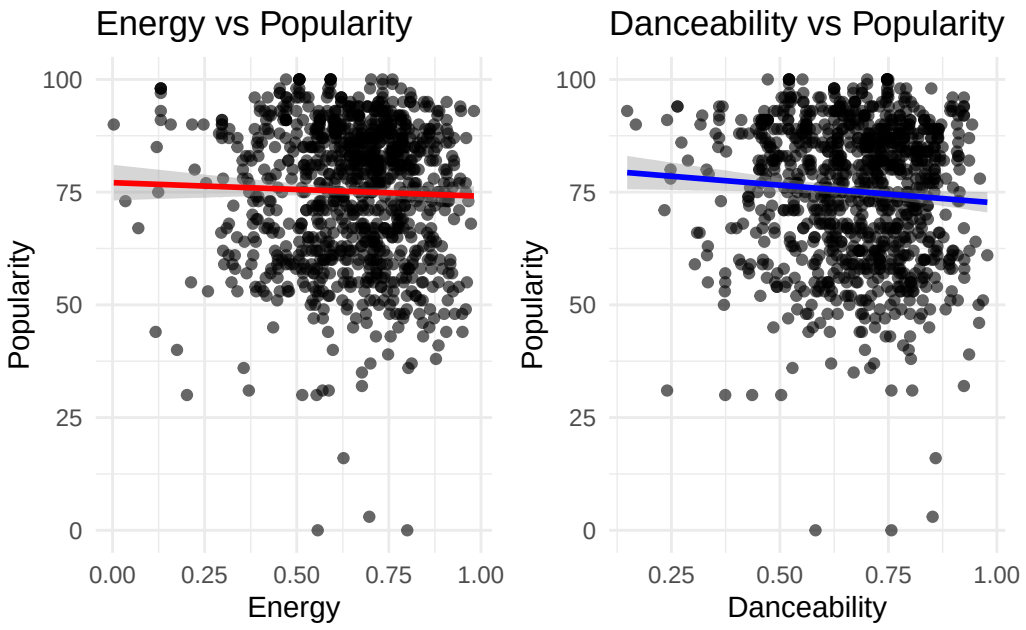
For our multiple regression model, we selected danceability as the second explanatory variable alongside energy for several reasons. First, both energy and danceability are fundamental characteristics that music industry professionals often consider when evaluating commercial potential; energy captures the intensity and power of a song, while danceability measures how suitable a track is for dancing based on tempo, rhythm stability, and beat strength. Second, these variables represent different but complementary aspects of musical appeal that could work together to influence popularity. Finally, energy and danceability are conceptually distinct enough that they should provide different information about song characteristics, while still being related enough to potentially interact in meaningful ways.

Data Visualization

We can use two visualization approaches to examine our multiple regression relationships. Our first approach includes individual scatterplots which show each of our predictor's separate relationship with popularity. Our second approach displays all three variables together using color coding, providing us with insight into how energy and danceability jointly relate to song popularity.

```
# energy vs popularity
p1 <- ggplot(spotify_sample, aes(x = energy, y = popularity)) +
  geom_point(alpha = 0.6) +
  geom_smooth(method = "lm", se = TRUE, color = "red") +
  labs(title = "Energy vs Popularity",
       x = "Energy",
       y = "Popularity") +
  theme_minimal()

p2 <- ggplot(spotify_sample, aes(x = danceability, y = popularity)) +
  geom_point(alpha = 0.6) +
  geom_smooth(method = "lm", se = TRUE, color = "blue") +
  labs(title = "Danceability vs Popularity",
       x = "Danceability",
       y = "Popularity") +
```

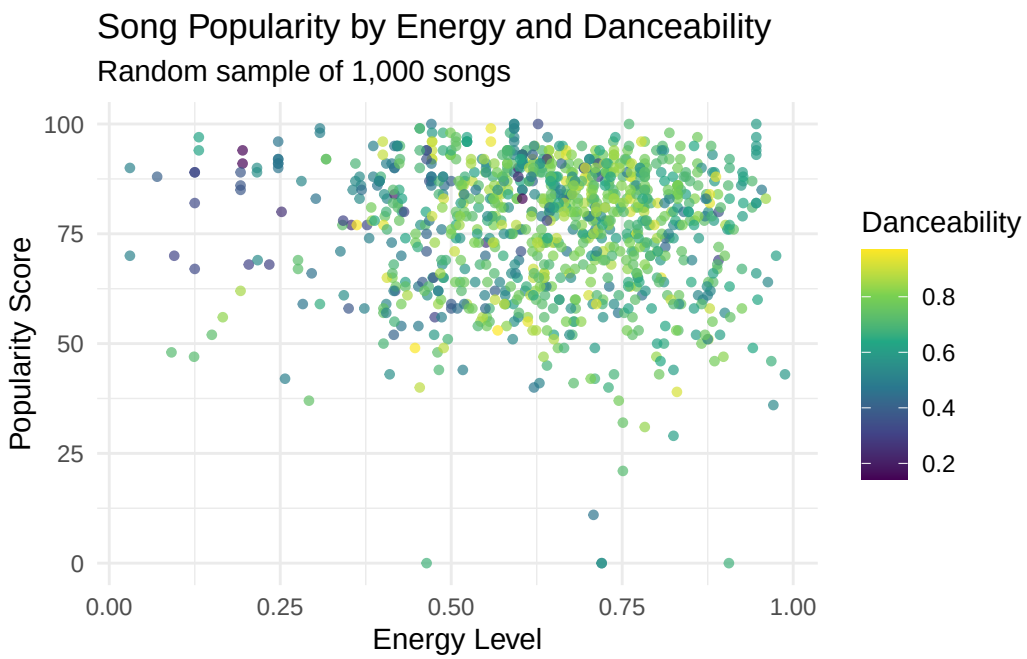
```
  theme_minimal()

# combining plots
library(patchwork)
p1 + p2
```



Earlier when we plotted energy level vs. popularity (also shown in the left), we found that there is a weak negative relationship between these two variables. The plot on the right also shows a weak negative relationship between danceability and popularity, with a similarly flat trend line. However, the danceability trend line appears to sit slightly higher than the energy trend line, suggesting that danceability may have a marginally more favorable baseline association with popularity compared to energy. Both relationships demonstrate considerable scatter around their respective trend lines, indicating that neither audio feature alone strongly predicts song popularity.

```
# taking a random sample of your data for plotting
spotify_sample <- spotify_data %>%
  slice_sample(n = 1000)

ggplot(spotify_sample, aes(x = energy, y = popularity, color = danceability)) +
  geom_point(alpha = 0.7, size = 1.2) +
  scale_color_viridis_c(name = "Danceability") +
  labs(title = "Song Popularity by Energy and Danceability",
```

```
      subtitle = "Random sample of 1,000 songs",
      x = "Energy Level",
      y = "Popularity Score") +
  theme_minimal()
```



Our combined visualization shows all three variables simultaneously, with energy on the x-axis, popularity on the y-axis, and danceability represented by color (yellow-green = high danceability, dark blue = low danceability). Our plot reveals that songs are distributed across the full range of energy and danceability combinations, with most songs clustering in the moderate-to-high energy range. Notably, there doesn't appear to be a strong clustering pattern based on danceability levels; both high-danceability songs (yellow-green) and low-danceability songs (dark blue) are found at similar popularity levels across different energy ranges. This suggests that the combination of energy and danceability may not create distinct patterns for predicting popularity, reinforcing that these audio features may have limited explanatory power for song success.

Model Fitting and Justification

Model Choice Justification

We will be using ordinary least squares (OLS) again to maintain consistency and ensure interpretability. Our multiple linear regression model will include both energy and danceability as explanatory variables to examine their combined effect on song popularity.

```
# fitting multiple linear regression model
model2 <- lm(popularity ~ energy + danceability, data = spotify_data)
summary(model2)
```

```
Call:
lm(formula = popularity ~ energy + danceability, data = spotify_data)

Residuals:
    Min      1Q  Median      3Q     Max
-79.296 -11.021   3.507  12.127  24.594

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  80.35829    0.05953 1349.82   <2e-16 ***
energy       -1.04051    0.06659  -15.62   <2e-16 ***
danceability -5.58657    0.07809  -71.54   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 15.74 on 2110313 degrees of freedom
Multiple R-squared:  0.003037,  Adjusted R-squared:  0.003036
F-statistic:  3214 on 2 and 2110313 DF,  p-value: < 2.2e-16
```

```
confint(model2)
```

```
                2.5 %     97.5 %
(Intercept)  80.241608 80.4749721
energy       -1.171032 -0.9099982
danceability -5.739614 -5.4335173
```

Our multiple linear regression results show statistically significant relationships for both predictors, though with limited practical impact. Our model estimates that for every 1-unit increase in energy level (holding danceability constant), popularity decreases by an average of 0.92 points (95% CI: -1.05 to -0.78, $p < 0.001$). More notably, for every 1-unit increase in danceability (holding energy constant), popularity decreases by an average of 5.38 points (95% CI: -5.54 to -5.23, $p < 0.001$), making danceability a stronger predictor than energy.

While both relationships are highly statistically significant due to the massive sample size, the combined model explains only 0.275% of the variation in popularity (R-squared = 0.00275). In other words, this means that even if we know both the energy and danceability of a song perfectly, we can only explain about one-quarter of one percent of why songs differ in popularity; the remaining 99.7% is determined by other factors. The intercept suggests that a hypothetical song with zero energy and zero danceability would have a popularity score of approximately 80, though this extrapolation falls outside the realistic range of the data. Despite the statistical significance, the extremely low R-squared value indicates that energy and danceability together provide minimal explanatory power for predicting song popularity, suggesting that other unmeasured factors play far more substantial roles in determining musical success.

Model Assumptions and Limitations

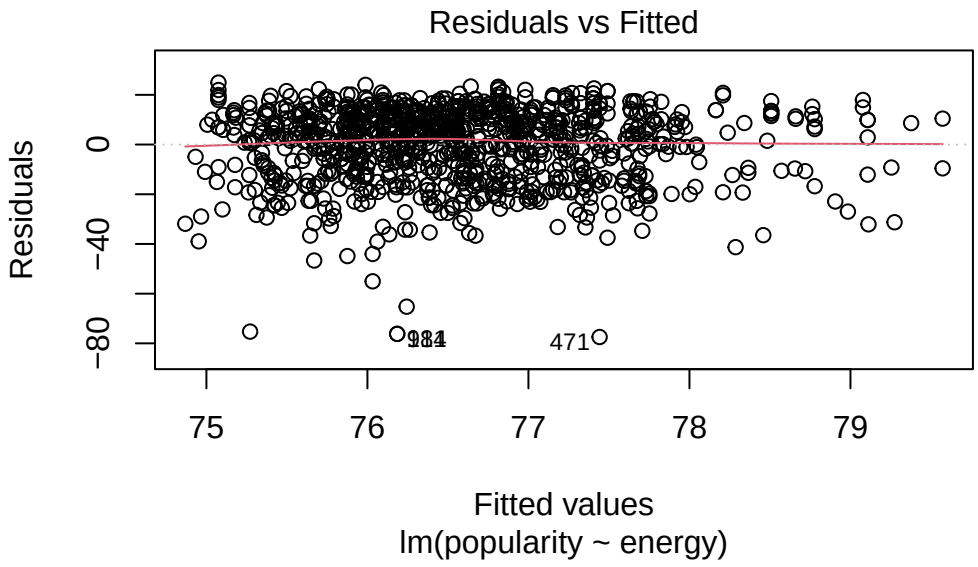
Linearity Assessment

Both models show reasonably linear relationships without strongly curved patterns, which makes linear regression appropriate for addressing our descriptive questions. While the relationships are weak, the scatterplots follow approximately straight-line trends with consistent variance, though some increased scatter occurs in certain ranges.

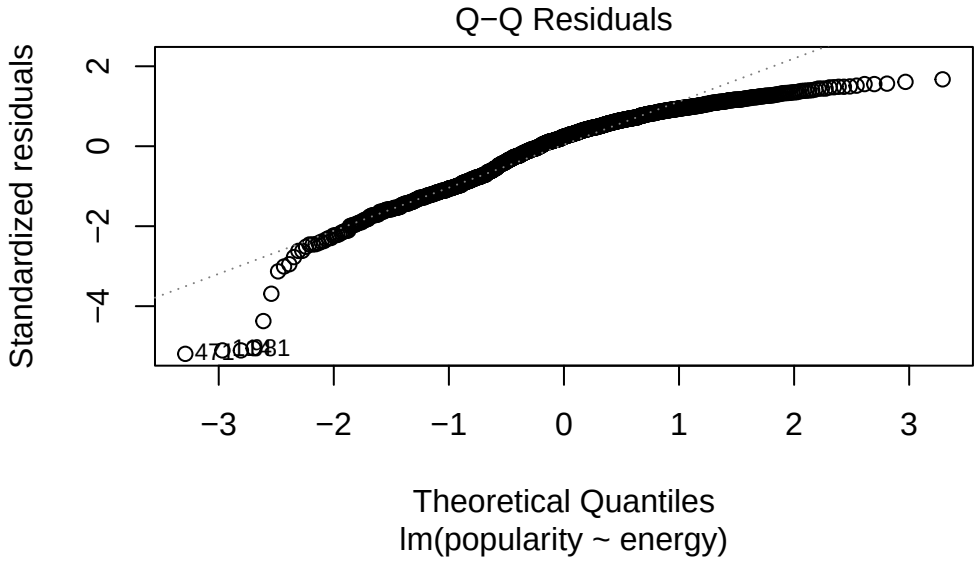
Residual Analysis

To test model assumptions a bit more thoroughly, we can examine residual plots:

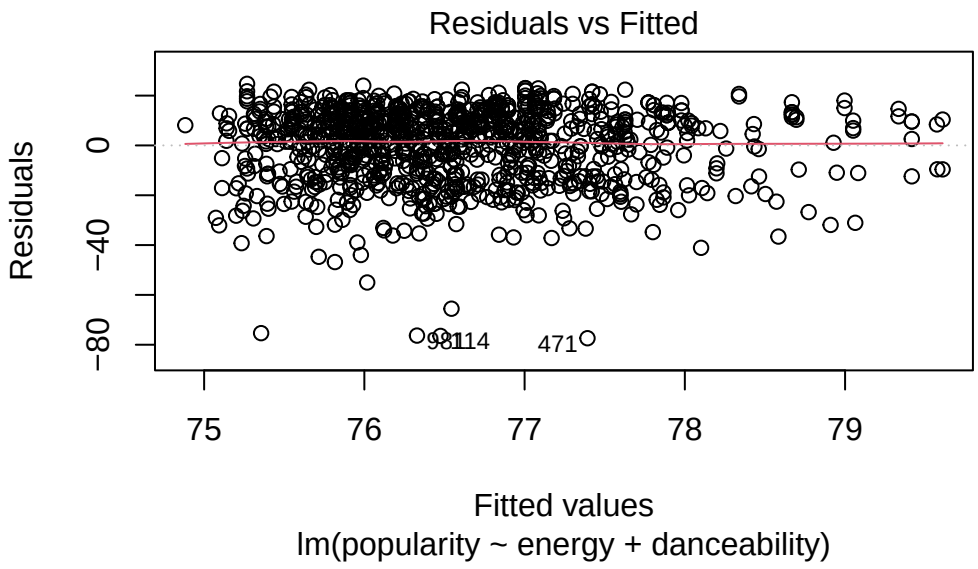
```
# for the simple linear regression model (model1)
model1_sample <- lm(popularity ~ energy, data = spotify_sample)
model2_sample <- lm(popularity ~ energy + danceability, data = spotify_sample)

# simple linear regression model diagnostics
plot(model1_sample, which = 1)  # residuals vs Fitted
```



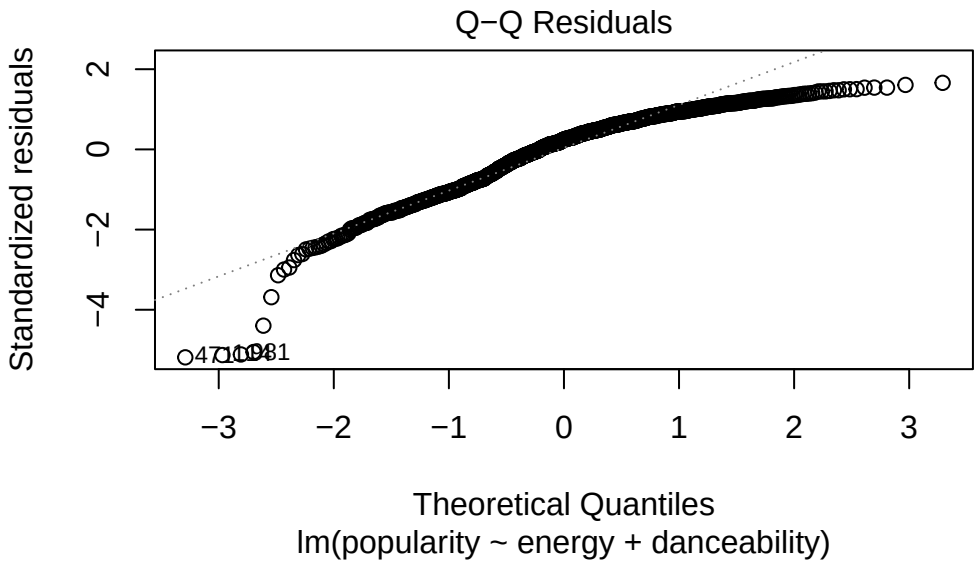

```r
plot(model1_sample, which = 2)  # Q-Q plot for normality
```



```r
# multiple linear regression model diagnostics
plot(model2_sample, which = 1)  # residuals vs Fitted
```




```r
plot(model2_sample, which = 2)  # Q-Q plot for normality
```



1. **Residuals vs. Fitted values**: Residuals vs. Fitted values: Show random scatter around zero with no clear patterns, confirming linearity and constant variance assumptions.
2. **Q-Q plots**: Residuals showed deviations from normality with heavy tails.

These plots largely confirm the appropriateness of linear modeling by visually checking for assumption violations like non-linearity, heteroscedasticity, or non-normality.

Model Appropriateness Assessment

While linear regression was technically appropriate given our data structure and research goals, the models' extremely low R-squared values suggest that audio features alone explain little about what drives song popularity.

Conclusion

Question 1 Results

Based on our simple linear regression analysis, energy has a statistically significant but practically negligible negative effect on popularity. Specifically, for every 1-unit increase in energy level, popularity decreases by an average of 2.15 points (95% CI: -2.28 to -2.02, $p < 0.001$). However, this relationship explains virtually none of the variation in song success, with energy accounting for only 0.053% of the variance in popularity. This finding suggests that while

there is a detectable relationship between energy and popularity, it is far too weak to be useful for practical prediction or decision-making.

Question 2 Results

Our multiple linear regression analysis shows that adding danceability provides minimal improvement in explaining song popularity. Energy decreases popularity by 0.92 points per unit increase (95% CI: -1.05 to -0.78, $p < 0.001$), while danceability has a stronger negative effect, decreasing popularity by 5.38 points per unit increase (95% CI: -5.54 to -5.23, $p < 0.001$). Together, both variables explain only 0.275% of the variation in popularity, making them ineffective predictors of song success despite their statistical significance. The model improvement from adding danceability is statistically significant but practically meaningless.

Validity Assessment

External Validity

This sample represents top songs from 73 countries, but may not generalize to all songs on Spotify or other music platforms, as it only includes already-popular tracks. To improve external validity, we should include songs that became viral through different pathways (social media trends, playlist features, organic discovery) rather than just chart-toppers, and examine songs across different time periods to account for changing trends in what drives popularity. Additionally, expanding to other platforms where songs go viral (TikTok, YouTube, etc.) would enhance generalizability of our findings about popular music characteristics.

Internal Validity

Since this is observational data, we cannot establish causation between audio features and popularity. Confounding variables like artist popularity, marketing, release timing, playlist placements, and/or social media presence could influence both song features and popularity. To improve internal validity, we should control for additional variables such as artist follower count, label size, marketing spend, and release strategy. Alternatively, examining naturally occurring experiments (such as comparing similar songs with different audio features from the same artist) could help isolate causal effects.

Construct Validity

Spotify's popularity measure may not capture all aspects of musical success, such as radio play, physical sales, cultural impact, or longevity. Additionally, algorithmically-determined song features like "energy" and "danceability" may not align with human perception of these characteristics. To improve construct validity, we should validate Spotify's popularity metric

against multiple success measures (chart positions, streaming revenue, social media engagement) and compare algorithmic audio features with human ratings of the same characteristics to ensure they measure what we intend.

Implications and Future Research

These findings, supported by the confidence intervals showing consistent negative effects with minimal explanatory power, suggest that music industry professionals should not rely solely on audio characteristics like energy or danceability when predicting song success. The extremely low predictive power indicates that marketing strategies, artist development, cultural positioning, and timing may be far more crucial investments than optimizing specific audio features. However, this doesn't mean audio features are irrelevant - they may play important roles in specific contexts or interact with unmeasured variables in complex ways.

Future directions for this analysis should incorporate variables like artist popularity, social media presence, playlist placements, release timing, and genre classifications to better understand the multifaceted nature of streaming success. Additionally, examining regional differences, temporal trends, platform-specific patterns, and the role of recommendation algorithms could provide deeper insights into what drives musical popularity in the digital age.

References

Saniczka, A. "Top Spotify Songs in 73 Countries (Daily Updated)." Kaggle, 2024. https://www.kaggle.com/datasets/asaniczka/top-spotify-songs-in-73-countries-daily-updated