Why We Dance: On Spotify R&B Artists

Short Report 1*

Dongyoung and Kunwu Lyu

Claire Kelling
STAT 230, Applied Regression Analysis
October 18, 2024

1 Introduction

People listen to music for a variety of reasons. Colley et al. (2022) looked into what descriptive features of the song contribute to its popularity. Their finding suggests that higher energy and danceability generally lead to higher appeals across platforms. Duman et al. (2022) investigated what factors lead people to listen to dance music. They found that dance music has, among other things, significantly higher levels of energy and loudness. Interested in how musical genres and racial boundaries were navigated during the 1960s, Brackett (1994) highlights how the process of crossing over reflects the racial tensions and segregation that characterized American society at the time. Does this change what music we dance to? In this study, we want to see if energy and loudness contribute to danceability in any way for specific R&B artists that we are interested in: Gladys Knight & The Pips, Stevie Wonder, and The Temptations,[1] and if the process of crossing over changes our interpretation of the model in any way.

2 Results

2.1 Data

We used data sourced from the Spotify API, a widely used music streaming platform, focusing on R&B music. We highlight the crossover variable here. It indicates whether a song ranked in the top 20 on both the Pop and R&B charts, meaning it was successful in both genres or "crossed over." See Appendix A for variable descriptions.

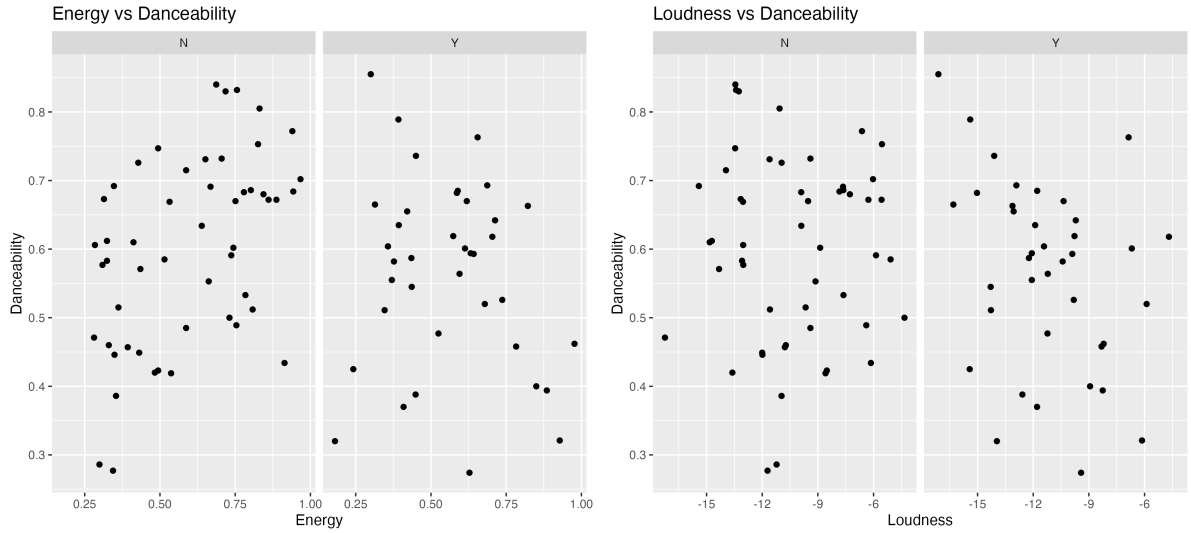
2.2 Exploratory Data Analysis

In order to explore the relationship between explanatory variables (Loudness, Energy, and Crossover category) and the response variable (Danceability), we generated scatterplots to visually check these relationships.

We plot the danceability of a song against its energy and loudness, respectively, in Figure 1, both separated by whether a song is a crossover hit. This allows us to investigate whether the Crossover Category affects the relationship between Energy and Danceability. From Figure 1a, while we do not see clear line trends, the relationship between Energy and Danceability seems more positive on non-crossover songs compared to crossover songs. We hypothesize that there is an association between energy and danceability. Furthermore, the effect of energy depends on if a song is a crossover hit.

*The source code of this project is available at my GitHub repository, including this LaTeX file.

[1]We used artists that our R&B class partners were assigned (Gladys Knight & The Pips from Kunwu's partner and Stevie Wonder & The Temptations from Dongyoung's partner).



(a) Danceability on Energy, by Crossover (b) Danceability on Loudness, by Crossover

Figure 1: Scatter Plot of Danceability on Energy and Loudness

Figure 1b is the scatter plots of Loudness versus Danceability, separated by Crossover Category (Y/N). On the two scatterplots, there appears to be some weak negative relationship between Loudness and Danceability. Unlike Figure 1a, however, we do not see a systematic difference between whether or not a song is a crossover hit. We, therefore, hypothesize that loudness affects danceability, to some extent.

2.3 Multiple Linear Regression

We give the proposed multiple regression model as follows:

$$\begin{aligned}\text{Danceability} = \beta_0 + \beta_1\,\text{Loudness} + \beta_2\,\text{Energy} + \beta_3\,\mathbb{1}(\text{Crossover} \;=\; \text{“Yes”}) + \\ \beta_4\,\mathbb{1}(\text{Crossover} \;=\; \text{“Yes”}) \times \text{Loudness} + \epsilon,\; \epsilon \overset{\text{i.i.d.}}{\sim} N(0,1). \quad (1)\end{aligned}$$

Our model shows good signs of overall fit ($F(4,\, 58) = 8.72$, $p < 0.001$, $R^2 = 0.29$).

2.4 Assumption Checking

We perform a residual analysis to verify our least squares assumptions. From Figure 2, we see that the residuals of both of our explanatory variables and the fitted values show signs of random scatter, indicating that the linearity and homoskedasticity assumptions are satisfied. From the normal $Q-Q$ plot, we see that the normality assumption is satisfied.[2]

As for the independence assumption, we grouped the residuals by artists and years to check for cluster and serial dependence. From Figure 3, we do not see evidence that the residuals depend on artists. Similar for years, the residuals show signs of random scatter. Therefore, our model assumptions are satisfied, and there is no need for transformations.

2.5 Interpretation

The results of our regression coefficients are given in Table 1. We have strong evidence that increasing loudness by one decibel is associated with a mean *decrease* of 0.02 in

[2]Though it is a bit heavy-tailed, our sample size ($N = 90$) ensures that this is not a concern.

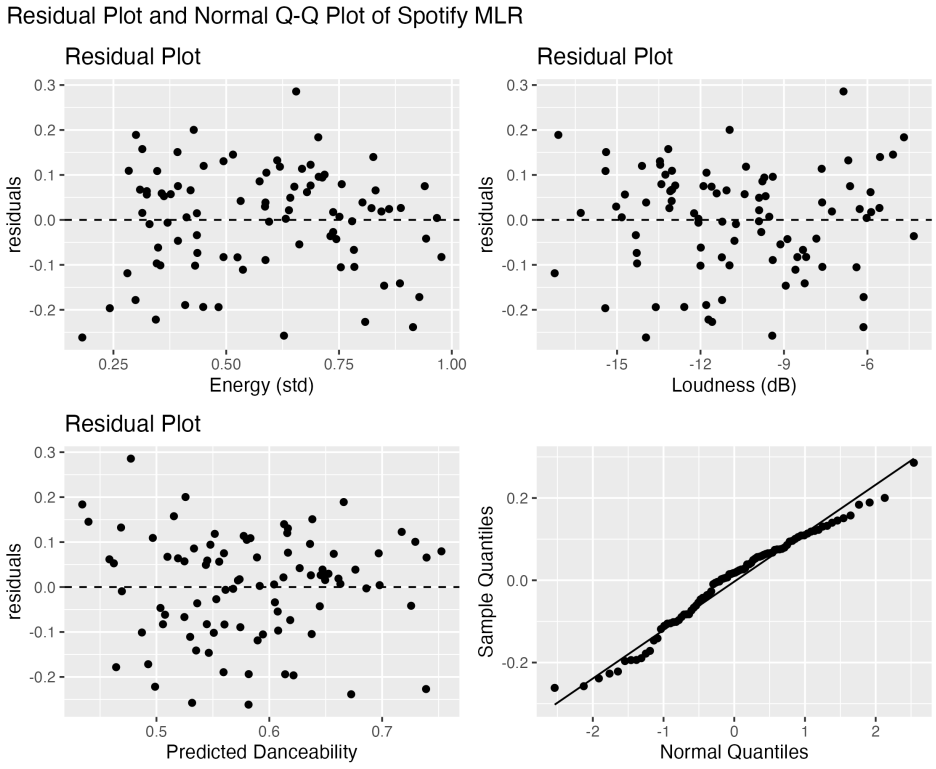


Figure 2: Residual and Normal $Q-Q$ Plots

danceability, after controlling for energy, whether or not a song is a crossover, and their interactions ($t = -4$, $p < 0.001$). Similarly for energy, there is strong evidence that a 0.1 unit *increase* in energy is associated with a mean increase of 0.052 units in danceability, controlling other variables ($t = 5.5$, $p < 0.001$). This is a very large effect, considering that danceability ranges from 0 to 1.

Interestingly, we also found moderately significant evidence that the effect of energy on danceability depends on whether or not this is a crossover song ($t = -3.2$, $p < 0.01$), confirming our conjecture earlier in subsection 2.2. A one-way ANOVA also reveals that the effect of energy does depend on whether the song is a crossover hit ($F = 6.67$, $d.f. = 2, 85$, $p = 0.002$). Specifically, the effect of energy on danceability increases when it is a crossover hit. There is also some evidence that fixing loudness and energy, whether or not this is a crossover song is associated with danceability ($t = 2.5$, $p < 0.05$).

	Estimate	Standard Error	t value	p-value (two sided)
Intercept	0.056	0.103	0.548	0.585
Loudness	-0.022	0.006	-4.025	$1 \times 10^{-4***}$
Energy	0.524	0.095	5.503	$3.88 \times 10^{-7***}$
Crossover (Y)	0.192	0.077	2.501	0.014^{*}
Energy * Crossover (Y)	-0.408	0.126	-3.245	0.002^{**}

Table 1: Regression Coefficients

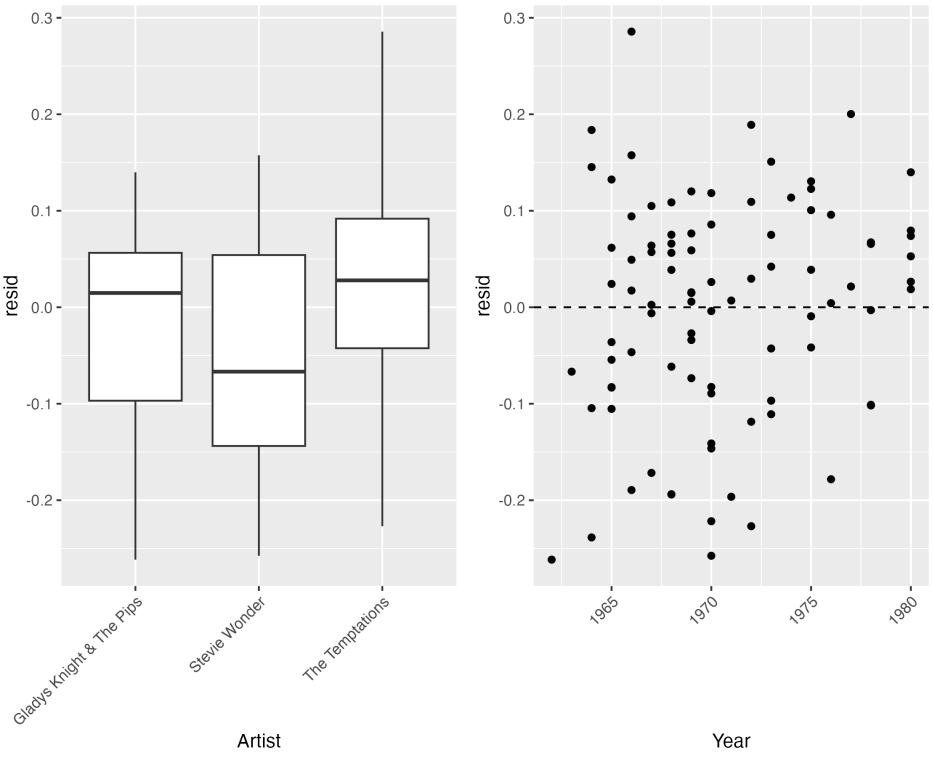


Figure 3: Independence Plots

3 Discussion

As we see from Appendix A, the concept of Energy contains some measurement of loudness. In fact, from Figure 4, we see that there is a strong, positive, and linear relationship between loudness and energy. It is, therefore, not surprising that they yielded significant results—they are, in some sense, dependent. While it is not within the scope of this study, it would be worthwhile to look into how those variables are defined so we can get a better grasp of our results. It would also be interesting to see how other variables affect danceability and whether they depend on crossover or not, such as tempo, mode, and so on.

Appendix A Variable Description

Danceability How well a track is suited for dancing, determined by various musical factors like tempo, rhythm stability, beat strength, and overall consistency. It is measured on a scale from 0.0 (least danceable) to 1.0 (most danceable).

Energy Rated between 0.0 and 1.0, representing the perceived intensity and activity of a song. High-energy tracks are often fast, loud, and dynamic. For instance, death metal has high energy, whereas a Bach prelude has low energy. Factors influencing this measure include dynamic range, perceived loudness, timbre, onset rate, and overall entropy.

Loudness Measured in decibels (dB), is the average volume level of a track throughout its duration. It helps in comparing the relative loudness between tracks. Loudness is related to the amplitude of sound and typically ranges from -60 to 0 dB.

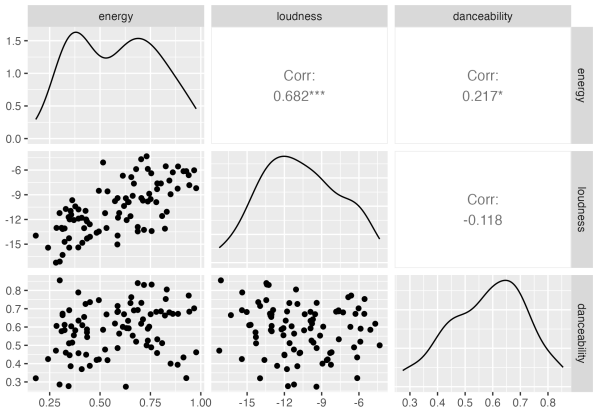


Figure 4: Scatterplot Matrix of the Data Set

Appendix B R Code Used

```
################################ DATA WRANGLING ###################################
#first spotify data, with crossover variable
#BE SURE TO DELETE THE LINE BREAK BEFORE RUNNING THIS
spotify_data_crossover <- read.csv(file =
"https://www.math.carleton.edu/ckelling/regression/report_cross_spotify_data.csv") %>%
  mutate(mode = as.factor(mode),
         key = as.factor(key),
         crossover_categ = as.factor(crossover_categ))

#second spotify dataset- all data, crossover labelled or not
#BE SURE TO DELETE THE LINE BREAK BEFORE RUNNING THIS
full_musc_data <- read.csv(file =
"https://www.math.carleton.edu/ckelling/regression/report_nocross_spotify_data.csv") %>%
  mutate(mode = as.factor(mode),
         key = as.factor(key))

# select specified artists and create a new data frame
```

```r
artists <- c("Gladys Knight & The Pips", "Stevie Wonder", "The Temptations")
spotify_data_crossover_filtered <- spotify_data_crossover %>%
  filter(artist_name %in%
           c("Gladys Knight & The Pips", "Stevie Wonder", "The Temptations"))

glimpse(spotify_data_crossover_filtered)

##################################### EDA #####################################
# scatterplot matrix
ggpairs(spotify_data_crossover_filtered,
        columns = c("energy", "loudness", "danceability"))
ggsave("scat_mtrx.png") # save as png for better importing

# Vis
# EDA with energy vs danceability
eda_energy <- ggplot(data = spotify_data_crossover_filtered,
                     aes(x = energy, y = danceability)) +
  geom_point() +
  facet_grid(cols = vars(crossover_categ)) + # separate by crossover
  labs(x = "Energy", y = "Danceability", title = "Energy vs Danceability")

# EDA with loudness vs danceability
eda_loudness <- ggplot(data = spotify_data_crossover_filtered,
                       aes(x = loudness, y = danceability)) +
  geom_point() +
  labs(x = "Loudness", y = "Danceability", title = "Loudness vs Danceability") +
  facet_grid(cols = vars(crossover_categ))  # separate by crossover

ggsave("eda_energy.png", eda_energy)
ggsave("eda_loudness.png", eda_loudness)

############################## MLR and ASSUMPTIONS ##############################
# MLR Model
danceability_lm <- lm(danceability ~ loudness + energy * crossover_categ,
                      data = spotify_data_crossover_filtered)
summary(danceability_lm)

# Assumptions Checking
danceability_lm_aug <- augment(danceability_lm, # add original data for later
                                                # checking assumptions
                               data = spotify_data_crossover_filtered)

# residual plot
danceability_lm_res1 <- ggplot(danceability_lm_aug, aes(x = energy, y = .resid)) +
  geom_point() +
  geom_hline(yintercept = 0, linetype = "dashed") +
  labs(x = "Energy (std)", y = "residuals",
       title = "Residual Plot")

danceability_lm_res2 <- ggplot(danceability_lm_aug, aes(x = loudness, y = .resid)) +
  geom_point() +
```

```
  geom_hline(yintercept = 0, linetype = "dashed") +
  labs(x = "Loudness (dB)", y = "residuals",
       title = "Residual Plot")

danceability_lm_res3 <- ggplot(danceability_lm_aug, aes(x = .fitted, y = .resid)) +
  geom_point() +
  geom_hline(yintercept = 0, linetype = "dashed") +
  labs(x = "Predicted Danceability", y = "residuals",
       title = "Residual Plot")

# normal qq plot
danceability_lm_qq <- ggplot(danceability_lm_aug, aes(sample = .resid))+
  geom_qq() +
  geom_qq_line() +
  labs(y = "Sample Quantiles", x = "Normal Quantiles")

combined_plot <- (danceability_lm_res1 | danceability_lm_res2) /
  (danceability_lm_res3 | danceability_lm_qq)
res_qq <- combined_plot +
  plot_layout(guides = 'collect') +
  plot_annotation(title = "Residual Plot and Normal Q-Q Plot of Spotify MLR")

res_qq <- ggsave("res+qq.png", res_qq)

# Cluster independence
cluster_dep <- ggplot(danceability_lm_aug, aes(x = artist_name, y = .resid)) +
  geom_boxplot() +
  labs(y = "resid", x = "Artist") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))

# Serial independence, w/ boxplot
ggplot(danceability_lm_aug, aes(x = as.factor(album_release_year), y = .resid)) +
  geom_boxplot() +
  labs(y = "resid", x = "Year") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))

# Serial independence, w/ scatterplot (better visuals)
serial_dep <- ggplot(danceability_lm_aug,
                     aes(x = album_release_year, y = .resid)) +
  geom_point() +
  geom_hline(yintercept = 0, linetype = "dashed") +
  # geom_smooth(method = "lm", se = T) +
  labs(y = "resid", x = "Year") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))

cluster_dep | serial_dep -> independence_check

ggsave("independence.png", independence_check)

##################################### R^2 #####################################
danceability_lm_red <- lm(danceability ~ loudness + energy,
```

```
data = spotify_data_crossover_filtered)

anova(danceability_lm_red, danceability_lm)
```

References

Brackett, D. (1994). The Politics and Practice of "Crossover" in American Popular Music, 1963 to 1965. *The Musical Quarterly*, 78(4):774–797. Publisher: Oxford University Press.

Colley, L., Dybka, A., Gauthier, A., Laboissonniere, J., Mougeot, A., Mowla, N., Dick, K., Khalil, H., and Wainer, G. (2022). Elucidation of the Relationship Between a Song's Spotify Descriptive Metrics and its Popularity on Various Platforms. In *2022 IEEE 46th Annual Computers, Software, and Applications Conference (COMPSAC)*, pages 241–249. ISSN: 0730-3157.

Duman, D., Neto, P., Mavrolampados, A., Toiviainen, P., and Luck, G. (2022). Music we move to: Spotify audio features and reasons for listening. *PLOS ONE*, 17(9):e0275228. Publisher: Public Library of Science.