# SDS 291 Final Project: Research Design

Olga Kazarov, Glenvelis Perez, and Sarah Mian

2023-11-21

**Background**

Today, music is a crucial part of social life, with people using it as a form of connection and emotional expression. Especially as platforms such as Spotify have become more popular, music has become more accessible. For this reason, we are focusing on the popularity of music, as measured by the number of streams a song receives on Spotify, because this platform is one of the most widely used music streaming services in the world right now. It is interesting to us to be able to see which factors are influencing the music that we are listening to and that the people around us are listening to as well, since the popularity of music is oftentimes an important factor in determining what songs we all come across and end up listening to. Our research will help further determine what factors actually play a role in song popularity.

We are interested in analyzing which factors play a role in the number of Spotify streams a song has. We will specifically be looking at artist features, the release time of the song, energy and valence, and beats per minute. We expect that songs with higher energy, higher valence, and a higher number of beats per minute will have a higher number of streams. Furthermore, we hypothesize that if a song is released in the summer, it will have higher energy levels and a higher number of beats per minute. Finally, we hypothesize that valence, energy, beats per minute, and whether a song was released in the summer will create the best predictor model for a song's number of streams.

Previous research has shown that the energy and valence of a song positively affect its popularity, while speechiness, instrumentalness, and liveness negatively affect popularity. It was also found that the interaction between energy and valence had a strong negative effect on popularity (Sciandra and Spera 2020). Studies have also shown seasonal changes in music preference. For example, the intensity of music is correlated with day length and thus there is an uptick in music intensity during the warmer summer and a decline to more calm music in the cold winter months, specifically during the late-December holidays (Park et al. 2019). The presence of a featured artist on a song has been shown to increase the song's popularity as well as the longevity of the song's popularity (Suh 2019).

**Data**

The dataset we are using is a collection of the most streamed songs on Spotify in 2023, collected directly from the Spotify streaming platform from Spotify users, and made accessible on Kaggle. There are 943 observations in this dataset. The sampled population consists of Spotify songs, and the sample was collected from Spotify in 2023. Our unit of analysis is songs. The variables that will be included in our analysis are energy percentage, valence percentage, number of beats per minute, number of streams, and whether or not a song was released in the summer; the audio feature variables were likely calculated through Spotify's own software.

**Variables**

We will transform the release month variable into a binary variable representing whether the month is a summer month (months 6-8) or is not a summer month (all remaining months). Similarly, we will transform the number of artists featured on a song to a binary variable (0 if no artists are featured and 1 if at least one artist is featured). Our primary dependent variable is the number of streams a song receives. Our qualitative independent variables are whether or not there are artist features on the song, and if the song was released in the summer; our quantitative independent variables are energy and valence percentages and beats per minute. We will use our summer binary variable as a dependent variable in one of our models to see if there is a relationship between release time and the audio features of a song.

**Method**

We will use our data to create two multiple regression models and one multiple logistic model. The first multiple regression model will focus on the correlation between valence percentage, energy percentage, and beats per minute with the number of streams. The second model is a multiple logistic model that will focus on energy percentage and beats per minute in relation to its probability of the song being released in the summer. The third model, our second multiple regression model, will use valence percentage, energy percentage, beats per minute, and whether a song was released in the summer to predict the number of streams.

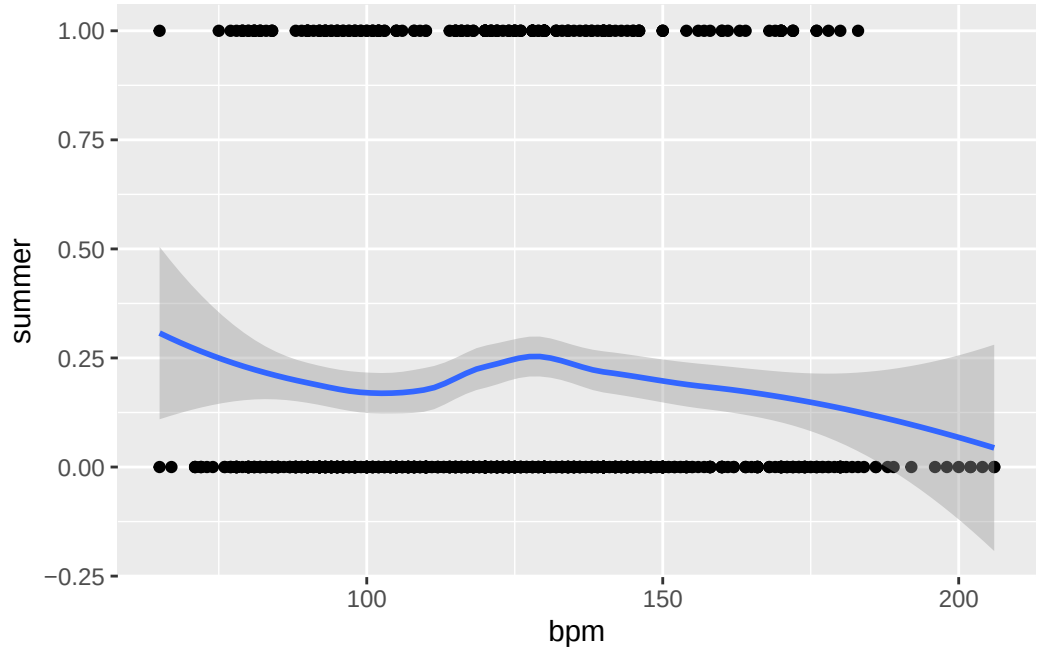
Our method will allow us to observe the best predictors in determining a song's number of streams as well as if there is a relationship between the timing of the song's release and its qualities, as well as its number of streams. Our approach allows us to take into account multiple variables and check for collinearity in our data.

```
spotify <- spotify[-c(124, 143, 145, 394), ]
```

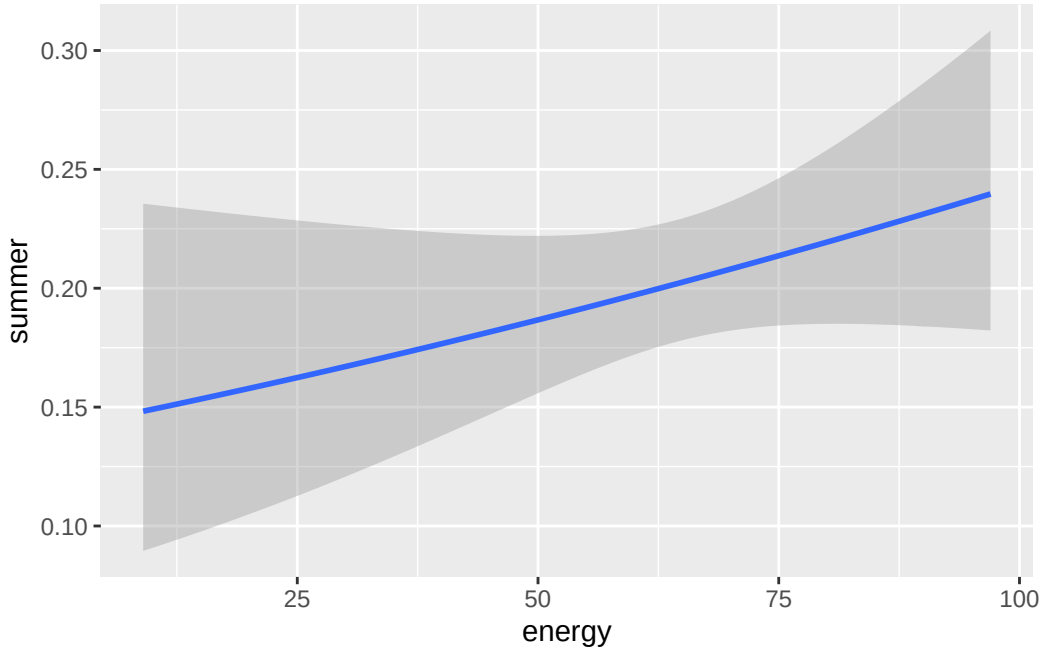
```
`geom_smooth()` using method = 'loess' and formula = 'y ~ x'
```

Table 1: Predicted Summer Song

| | *Dependent variable:* |
|---|---|
| | Summer |
| Beats Per Minute | −0.003 |
| | (0.003) |
| Energy Percentage | 0.007 |
| | (0.005) |
| Constant | −1.477*** |
| | (0.480) |
| Observations | 949 |
| Log Likelihood | −476.520 |
| Akaike Inf. Crit. | 959.040 |
| *Note:* | *p<0.1; **p<0.05; ***p<0.01 |




```
`geom_smooth()` using formula = 'y ~ x'
```




```
names(spotify)[names(spotify) == "valence_%"] <- "valence"
names(spotify)[names(spotify) == "danceability_%"] <- "danceability"
names(spotify)[names(spotify) == "acousticness_%"] <- "acousticness"
names(spotify)[names(spotify) == "instrumentalness_%"] <- "instrumentalness"
names(spotify)[names(spotify) == "liveness_%"] <- "liveness"
names(spotify)[names(spotify) == "speechiness_%"] <- "speechiness"
```

```
spotify <- subset(spotify, select = -c(released_year, released_day, in_spotify_playlists,
```

```
spotify <- spotify[-c(575), ]
spotify <- transform(spotify, streams = as.numeric(streams))
```

```
Warning in eval(substitute(list(...)), `_data`, parent.frame()): NAs introduced
by coercion
```

```
spotify <- spotify %>%
  mutate(streams = log(streams))
```

```r
out <- boxplot.stats(spotify$streams)$out
out_ind <- which(spotify$streams %in% c(out))
out_ind
```

```
integer(0)
```

```r
spotify[out_ind, ]
```

```
 [1] track_name     artist.s._name artist_count   released_month streams
 [6] bpm            valence        energy         features       summer
<0 rows> (or 0-length row.names)
```

```r
spotify <- spotify[-c(124, 143, 145, 394), ]
```

```r
model2 <- lm(streams~bpm+energy+valence+as.factor(summer)+as.factor(features), data=spotif
summary(model2)
```

```
Call:
lm(formula = streams ~ bpm + energy + valence + as.factor(summer) +
    as.factor(features), data = spotify)

Residuals:
    Min      1Q  Median      3Q     Max
-3.0078 -0.7372 -0.0400  0.8102  2.4534

Coefficients:
                       Estimate Std. Error t value Pr(>|t|)
(Intercept)          19.6168233  0.2035269  96.384  < 2e-16 ***
bpm                   0.0004285  0.0012290   0.349 0.727435
energy                0.0003872  0.0022375   0.173 0.862637
valence              -0.0004788  0.0015659  -0.306 0.759872
as.factor(summer)1   -0.1654658  0.0855190  -1.935 0.053310 .
as.factor(features)1 -0.2804703  0.0718435  -3.904 0.000101 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.053 on 937 degrees of freedom
```

```
  (1 observation deleted due to missingness)
Multiple R-squared:  0.02145,   Adjusted R-squared:  0.01623
F-statistic: 4.108 on 5 and 937 DF,  p-value: 0.001075
```

```
#model2_resid <- augment(model2)
#ggplot(model2, aes(x=.fitted, y=.resid)) + geom_point() + geom_smooth(method=lm)
```

```
stargazer(model2, type = "latex",
          title            = "Predicted Number of Streams",
          covariate.labels = c("BPM", "Energy", "Valence", "Summer", "Features"),
          dep.var.labels   = ("Streams"), header=F)
```

## Bibliography


Park, Minsu, Jennifer Thom, Sarah Mennicken, Henriette Cramer, and Michael Macy. 2019. "Global Music Streaming Data Reveal Diurnal and Seasonal Patterns of Affective Preference." *Nature Human Behaviour* 3 (3): 230–36. https://doi.org/10.1038/s41562-018-0508-z.

Sciandra, Mariangela, and Irene Carola Spera. 2020. "A Model-Based Approach to Spotify Data Analysis: A Beta GLMM." *Journal of Applied Statistics* 49 (1): 214–29. https://doi.org/10.1080/02664763.2020.1803810.

Suh, Brendan. 2019. "International Music Preferences: An Analysis of the Determinants of Song Popularity on Spotify for the U.S., Norway, Taiwan, Ecuador, and Costa Rica." *CMC Senior Theses*, January. https://scholarship.claremont.edu/cmc_theses/2271.

Table 2: Predicted Number of Streams

| | *Dependent variable:* |
|---|---|
| | Streams |
| BPM | 0.0004 |
| | (0.001) |
| Energy | 0.0004 |
| | (0.002) |
| Valence | −0.0005 |
| | (0.002) |
| Summer | −0.165* |
| | (0.086) |
| Features | −0.280*** |
| | (0.072) |
| Constant | 19.617*** |
| | (0.204) |
| Observations | 943 |
| $R^2$ | 0.021 |
| Adjusted $R^2$ | 0.016 |
| Residual Std. Error | 1.053 (df = 937) |
| F Statistic | 4.108*** (df = 5; 937) |
| *Note:* | *p<0.1; **p<0.05; ***p<0.01 |