Spotify Analysis

Predicting danceability using the most streamed Spotify songs

By Amar, Elliott, and Trung

Table of Contents

Summary of Data

Obtained from Kaggle, this dataset contains the **most streamed songs of 2023 on the music platform Spotify**. The data includes information about each song's characteristics and its popularity on various music platforms.

We analyzed how these musical factors impact a song's "danceability."

track_name	released_...	in_spotify...	streams	in_apple_...	bpm	danceabili...	valence_%	energy_%	acousticn...	instrumen...	speechin...
Seven (feat. Latto) (Explicit Ver.)	7	553	141381703	43	125	80	89	83	31	0	4
LALA	3	1474	133716286	48	92	71	61	74	7	0	4
vampire	6	1397	140003974	94	138	51	32	53	17	0	6
Cruel Summer	8	7858	800840817	116	170	55	58	72	11	0	15
WHERE SHE GOES	5	3133	303236322	84	144	65	23	80	14	63	6
Sprinter	6	2186	183706234	67	141	92	66	58	19	0	24

Variable Overview

Dependent Variable:
danceability_%: *percentage indicating how suitable the song is for dancing*

Independent Variables:
bpm: *beats per minute, a measure of song tempo*
valence_%: *positivity of the song's musical content*
energy_%: *perceived energy level of the song*
acousticness_%: *amount of acoustic sound in the song*
speechiness_%: *amount of spoken words in the song*
liveness_%: *presence of live performance elements*
instrumentalness_%: *amount of instrumental content in the song*

Stepwise Regression

Forward Regression:

```
Step:  AIC=4817.03
danceability ~ valence + acousticness + bpm + speechiness + liveness +
    energy

                   Df Sum of Sq    RSS    AIC
<none>                          147195 4817.0
+ instrumentalness  1    53.799 147142 4818.7

Call:
lm(formula = danceability ~ valence + acousticness + bpm + speechiness +
    liveness + energy, data = spotify_data)

Coefficients:
 (Intercept)       valence  acousticness           bpm   speechiness      liveness        energy
    72.33862       0.26134      -0.14250      -0.09038       0.24577      -0.09159      -0.07314
```

Backward Regression:

```
Step:  AIC=4817.03
danceability ~ bpm + valence + energy + acousticness + liveness +
    speechiness

               Df Sum of Sq    RSS    AIC
<none>                     147195 4817.0
- energy        1     783.9 147979 4820.1
- liveness      1    1478.4 148674 4824.6
- speechiness   1    5619.8 152815 4850.7
- bpm           1    6101.2 153297 4853.7
- acousticness  1    8458.8 155654 4868.3
- valence       1   30295.8 177491 4993.4

Call:
lm(formula = danceability ~ bpm + valence + energy + acousticness +
    liveness + speechiness, data = spotify_data)

Coefficients:
 (Intercept)           bpm       valence        energy  acousticness      liveness   speechiness
    72.33862      -0.09038       0.26134      -0.07314      -0.14250      -0.09159       0.24577
```

Both the forward and backward stepwise regressions eliminated “instrumentalness” for variable selection.

They both support using bpm, valence, energy, acousticness, liveness, and speechiness to predict danceability.

First order model

```
Call:
lm(formula = danceability ~ valence + acousticness + bpm + speechiness +
    liveness + energy, data = spotify_data)

Residuals:
    Min      1Q  Median      3Q     Max
-38.536  -8.712   1.456   9.108  28.202

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  72.33862    2.92304  24.748  < 2e-16 ***
valence       0.26134    0.01873  13.954  < 2e-16 ***
acousticness -0.14250    0.01933  -7.373 3.64e-13 ***
bpm          -0.09038    0.01443  -6.262 5.76e-10 ***
speechiness   0.24577    0.04089   6.010 2.65e-09 ***
liveness     -0.09159    0.02971  -3.082  0.00211 **
energy       -0.07314    0.03258  -2.245  0.02502 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 12.47 on 946 degrees of freedom
Multiple R-squared:  0.2777,    Adjusted R-squared:  0.2731
F-statistic: 60.61 on 6 and 946 DF,  p-value: < 2.2e-16
```

After using stepwise regression to help pick our variables, we made a complete first order model using bpm, valence, energy, acousticness, liveness, and speechiness to predict danceability.

Adjusted $R^2 = 0.2731$
p-value = 2.2e-16 < $\alpha = 0.05$

```
> vif(model1)
     valence acousticness          bpm  speechiness     liveness       energy
    1.183279     1.544506     1.003345     1.005482     1.015541     1.779338
```

Second order model

We then made a complete second order model to capture more complex relationships between variables.

Adjusted $R^2 = 0.4244$
p-value = 2.2e-16 < α = 0.05

We noticed that many variables were not statistically significant. For example, acousticness and liveness were no longer statistically significant on their own as main effects, suggesting their prediction of danceability may be explained by interaction terms.

```
Residuals:
    Min      1Q  Median      3Q     Max
-33.496  -6.774   1.261   7.653  25.686

Coefficients:
                           Estimate Std. Error t value Pr(>|t|)
(Intercept)              -1.744e+01  1.312e+01  -1.329 0.184128
valence                   3.005e-01  1.354e-01   2.220 0.026659 *
acousticness             -2.182e-01  1.484e-01  -1.471 0.141711
bpm                       1.058e+00  1.231e-01   8.593  < 2e-16 ***
speechiness               1.411e+00  3.088e-01   4.571 5.52e-06 ***
liveness                 -2.180e-01  1.661e-01  -1.313 0.189645
energy                    5.486e-01  2.633e-01   2.084 0.037465 *
I(valence^2)             -3.229e-04  7.263e-04  -0.445 0.656688
I(acousticness^2)        -9.124e-04  8.139e-04  -1.121 0.262571
I(bpm^2)                 -4.562e-03  4.038e-04 -11.298  < 2e-16 ***
I(speechiness^2)         -1.950e-02  3.093e-03  -6.306 4.43e-10 ***
I(liveness^2)             1.121e-03  1.236e-03   0.907 0.364570
I(energy^2)              -3.459e-03  1.800e-03  -1.921 0.055011 .
valence:acousticness      4.798e-04  8.566e-04   0.560 0.575487
valence:bpm              -9.323e-05  6.023e-04  -0.155 0.877026
valence:speechiness      -1.720e-03  2.015e-03  -0.853 0.393661
valence:liveness          1.141e-03  1.405e-03   0.812 0.416914
valence:energy           -5.623e-04  1.443e-03  -0.390 0.696907
acousticness:bpm          1.072e-03  6.134e-04   1.748 0.080745 .
acousticness:speechiness -4.575e-03  1.646e-03  -2.780 0.005553 **
acousticness:liveness    -3.208e-03  1.306e-03  -2.457 0.014199 *
acousticness:energy       2.483e-03  1.631e-03   1.522 0.128371
bpm:speechiness           4.040e-03  1.228e-03   3.289 0.001043 **
bpm:liveness              8.319e-04  1.013e-03   0.821 0.411711
bpm:energy               -1.222e-03  1.045e-03  -1.169 0.242605
speechiness:liveness      1.778e-03  2.859e-03   0.622 0.534218
speechiness:energy       -1.094e-02  2.810e-03  -3.893 0.000106 ***
liveness:energy          -5.626e-04  2.255e-03  -0.249 0.803058
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 11.1 on 925 degrees of freedom
Multiple R-squared:  0.4407,    Adjusted R-squared:  0.4244
F-statistic: 26.99 on 27 and 925 DF,  p-value: < 2.2e-16
```

ANOVA

```
Analysis of Variance Table

Model 1: danceability ~ valence + acousticness + bpm + speechiness + liveness +
    energy
Model 2: danceability ~ valence + acousticness + bpm + speechiness + liveness +
    energy + valence * acousticness + valence * bpm + valence *
    speechiness + valence * liveness + valence * energy + acousticness *
    bpm + acousticness * speechiness + acousticness * liveness +
    acousticness * energy + bpm * speechiness + bpm * liveness +
    bpm * energy + speechiness * liveness + speechiness * energy +
    liveness * energy + I(valence^2) + I(acousticness^2) + I(bpm^2) +
    I(speechiness^2) + I(liveness^2) + I(energy^2)
  Res.Df    RSS Df Sum of Sq      F    Pr(>F)
1    946 147195
2    925 113979 21     33216 12.837 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

We performed an ANOVA test to ensure the inclusion of interaction and quadratic terms contributed to the prediction of danceability.

Ho: $\beta_{\text{interaction terms}} + \beta_{\text{quadratic terms}} = 0$
Ha: At least 1 $\beta_{\text{interaction term}}$ or $\beta_{\text{quadratic term}} \neq 0$

p-value = 2.2e-16 < $\alpha = 0.05$ so we reject Ho and conclude the second order model is **more statistically useful** than the first for predicting danceability.

Reduced model

```
Call:
lm(formula = danceability ~ valence + acousticness + bpm + speechiness +
    energy + I(bpm^2) + I(speechiness^2) + acousticness * speechiness +
    bpm * speechiness + speechiness * energy, data = spotify_data)

Residuals:
    Min      1Q  Median      3Q     Max
-34.586  -7.273   1.359   8.343  25.771

Coefficients:
                            Estimate Std. Error t value Pr(>|t|)
(Intercept)               -6.2904162  7.2824200  -0.864  0.38793
valence                    0.2492456  0.0173394  14.375  < 2e-16 ***
acousticness              -0.0709529  0.0249954  -2.839  0.00463 **
bpm                        1.0264970  0.1065639   9.633  < 2e-16 ***
speechiness                1.6174155  0.3133197   5.162 2.98e-07 ***
energy                     0.0217485  0.0424865   0.512  0.60885
I(bpm^2)                  -0.0045492  0.0004082 -11.144  < 2e-16 ***
I(speechiness^2)          -0.0209068  0.0031662  -6.603 6.71e-11 ***
acousticness:speechiness  -0.0043059  0.0016095  -2.675  0.00760 **
bpm:speechiness            0.0035655  0.0012506   2.851  0.00445 **
speechiness:energy        -0.0129654  0.0027296  -4.750 2.35e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 11.47 on 942 degrees of freedom
Multiple R-squared:  0.392,     Adjusted R-squared:  0.3856
F-statistic: 60.74 on 10 and 942 DF,  p-value: < 2.2e-16
```

After removing all the statistically insignificant variables from our second order model, we are left with the following reduced model.

While Adjusted R^2 decreased slightly (0.4244 to 0.3856), the model is now **much simpler** (27 vs 10 βs).

In this reduced model, energy became statistically insignificant as a main effect, likely because its effect on danceability is better captured through its interaction with speechiness.

Reduced model

$$\hat{y} = -6.290 + 0.249x_1 - 0.071x_2 + 1.026x_3 + 1.617x_4 + 0.022x_5 - 0.005x_3^2 - 0.021x_4^2 - 0.004x_2x_4 + 0.004x_3x_4 - 0.013x_4x_5$$

Where
x_1 = valence
x_2 = acousticness
x_3 = bpm
x_4 = speechiness
x_5 = energy

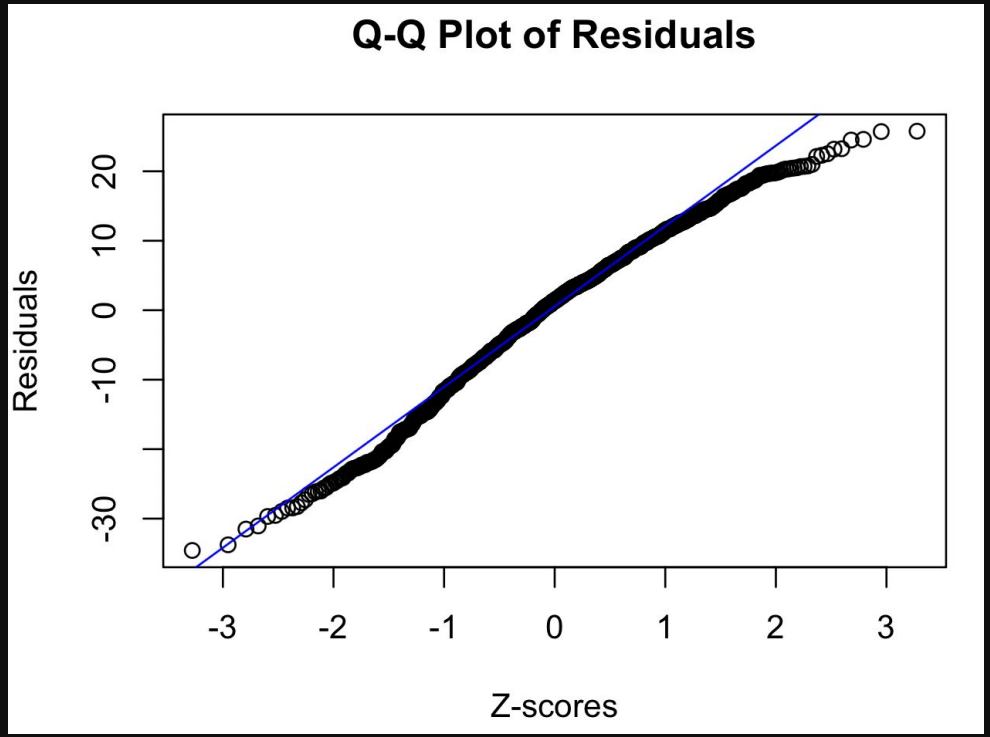
~38.56% of the variation in danceability (y) can be explained by the model after adjusting for the number of independent variables.

Residuals

The validity of many of the inferences associated with our regression analysis depends on the error term. We must check our assumptions that:

#		Assumptions		
1	♡	ε is normally distributed,	2:36	
2	♡	with a mean of 0,	3:21	
3	♡	the variance of σ^2 is constant,	3:04	
4	♡	all pairs of error terms are uncorrelated (independent)	2:47	

Checking that ε is normally distributed



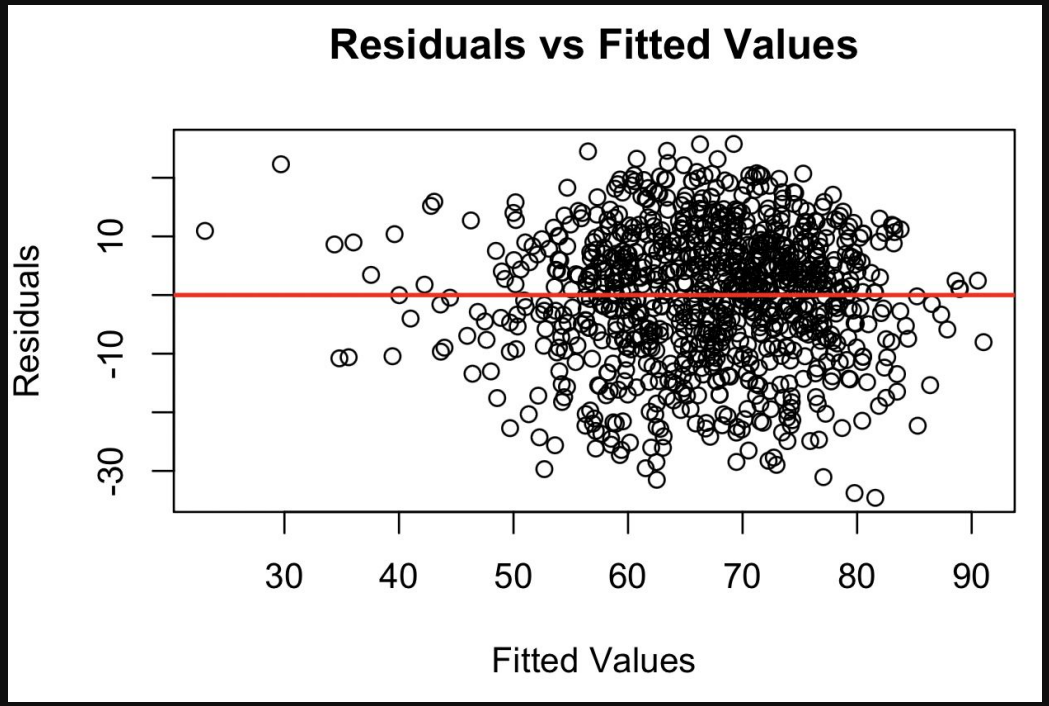
The Q-Q Plot is essentially a straight line, confirming that the residuals are normal.

Checking that ε has mean 0

```
> mean_residual <- mean(reduced_model$residuals)
> print(mean_residual)
[1] -2.332284e-16
```

The mean is very close to 0, so this assumption is satisfied.

Checking that the variance of σ^2 is constant



The residuals fall randomly around 0 without a clear pattern or systematic spread, suggesting constant (homoscedastic) variance σ^2 .

Examples

Song Name	Artist(s)	Valence	Acousticness	BPM	Speechiness	Energy	Danceability	Model Prediction
We Found Love	Rihanna, Calvin Harris	60	3	128	4	77	73	70.899
deja vu	Olivia Rodrigo	22	61	181	9	60	44	42.234
Locked Out of Heaven	Bruno Mars	87	6	144	5	70	73	75.438

```
> song_list <- c("We Found Love", "deja vu", "Locked Out Of Heaven")
> predicted_results <- predict_danceability(song_list, reduced_model, spotify_data_subset)
> print(predicted_results)
                 Song Predicted_Danceability
1        We Found Love               70.89874
2              deja vu               42.23446
3 Locked Out Of Heaven               75.43778
```

Thanks !