
What Makes Music Good?

Data

- Collected from the Spotify Web API by Kaggle user, tomigelo
- Updated April 2019
- Over 130,000 unique songs
- Each song is identified using three variables

Artist Name	Track ID	Track Name
Ariana Grande	1jQi0IyQL0ZsUinjty43nk	The Way - Spanglish Version
Free Throw	4Qx8LMAsSNXCrTBJjObp9F	Andy and I, Uh... (Revisited)
Adam Hawley	5S6EHC45If9NSN5jIHVzP1	Just Dance (feat. Dave Koz)
Brock Berrigan	0Ko2vMOAPx2rdJFFaYsuWA	Lake Powell
Miguel Y Miguel	3ki4CDv5x4MvT6qHSZx2kO	Cuando Ya No Me Quieras

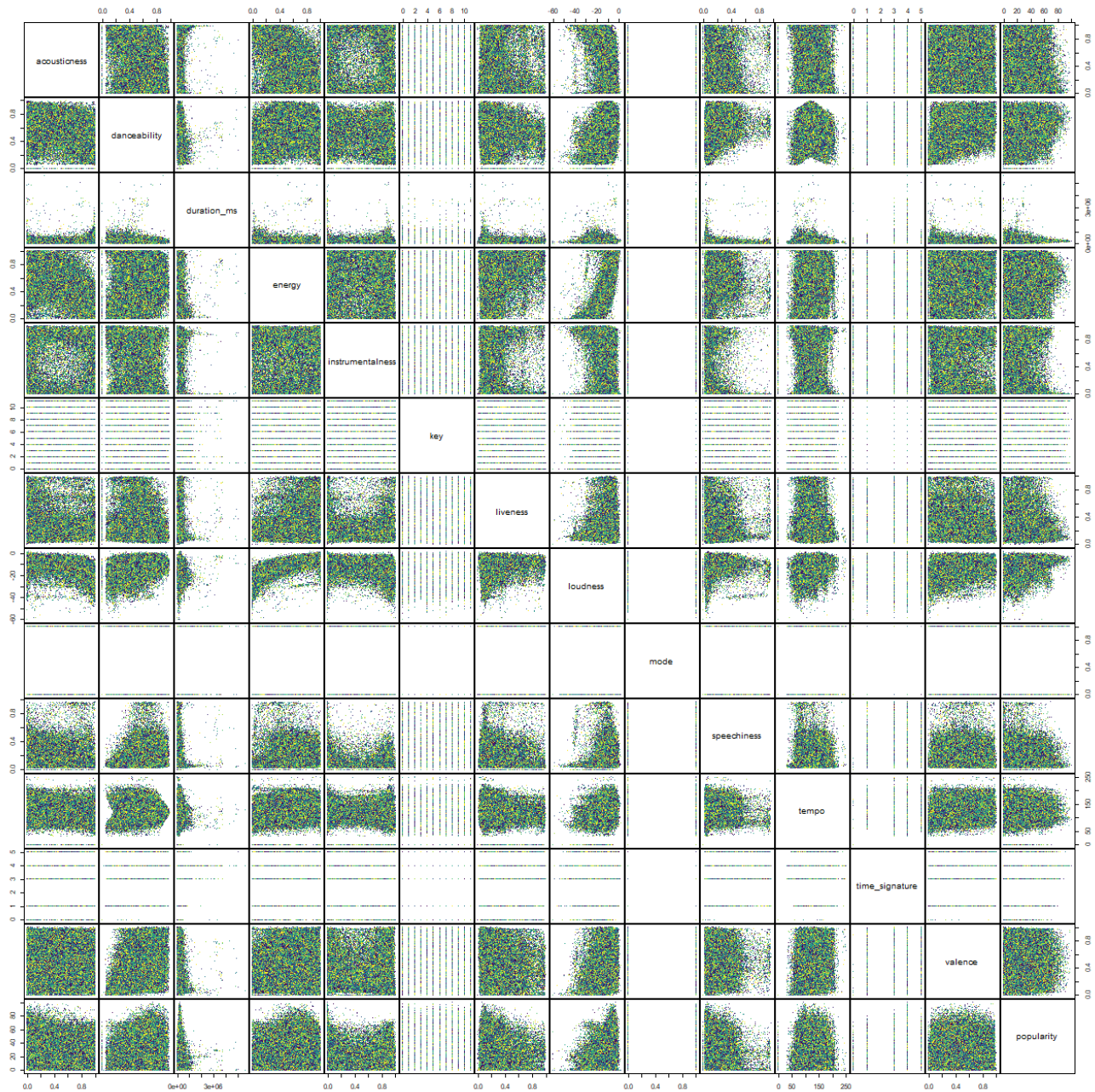
Feature	Min	Max	Median	Mean
Acousticness	0.0000	0.9960	0.2030	0.3425
Danceability	0.0000	0.9960	0.6050	0.5815
Duration_Ms	3203.0000	5610020.0000	201901.0000	212633.0000
Energy	0.0000	1.0000	0.6030	0.5692
Instrumentalness	0.0000	1.0000	0.0001	0.2240
Key	0.0000	11.0000	5.0000	5.2320
Liveness	0.0000	0.9990	0.1240	0.1949
Loudness	-60.0000	1.8060	-7.9790	-9.9740
Mode	0.0000	1.0000	1.0000	0.6077
Speechiness	0.0000	0.9660	0.0559	0.1120
Tempo	0.0000	249.9800	120.0300	119.4700
Time_Signature	0.0000	5.0000	4.0000	3.8790
Valence	0.0000	1.0000	0.4200	0.4396
Popularity	0.0000	100.0000	22.0000	24.2100

Data

- Every song has 13 audio features and one popularity feature

Matrix Plot

- Categorical variables included
- Not many relationships
- A little quadratic relationship with loudness and energy



Correlation to Popularity

ACOUSTICNESS	DANCEABILITY	DURATION MS	ENERGY	INSTRUMENTALNESS	KEY	LIVENESS	LOUDNESS	MODE	SPEECHINESS	TEMPO	TIME SIGNATURE	VALENCE
-0.1165	0.1311	-0.0008	0.1225	-0.2164	0.0027	-0.0312	0.2441	-0.0091	-0.0002	0.0371	0.0649	0.0143

- Very low correlation on all variables
- Strongest positive relationships: loudness, danceability, energy
- Strongest negative relationships: instrumentalness, acousticness, liveness

Linear Models

Full

```
Call:
lm(formula = popularity ~ ., data = spotifyCSVnums)

Residuals:
    Min      1Q  Median      3Q     Max
-35.868 -15.095  -1.858  12.737  74.193

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)       3.584e+01  6.334e-01  56.580  < 2e-16 ***
acousticness      7.491e-01  2.253e-01   3.324 0.000887 ***
danceability      5.451e+00  3.622e-01  15.049  < 2e-16 ***
duration_ms      -8.402e-07  4.308e-07  -1.951 0.051118 .
energy           -6.612e+00  3.880e-01 -17.039  < 2e-16 ***
instrumentalness -7.003e+00  1.762e-01 -39.741  < 2e-16 ***
key              -8.465e-03  1.472e-02  -0.575 0.565146
liveness         -2.878e+00  3.308e-01  -8.700  < 2e-16 ***
loudness          7.896e-01  1.485e-02  53.180  < 2e-16 ***
mode             -1.717e-01  1.092e-01  -1.573 0.115796
speechiness      -5.963e+00  4.499e-01 -13.255  < 2e-16 ***
tempo            -2.704e-03  1.793e-03  -1.508 0.131587
time_signature    8.088e-01  1.046e-01   7.729 1.09e-14 ***
valence          -7.067e+00  2.397e-01 -29.481  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 18.84 on 130649 degrees of freedom
Multiple R-squared:  0.08705,   Adjusted R-squared:  0.08696
F-statistic: 958.3 on 13 and 130649 DF,  p-value: < 2.2e-16
```

Reduced (Insignificant variables removed)

```
Call:
lm(formula = popularity ~ ., data = lm2)

Residuals:
    Min      1Q  Median      3Q     Max
-35.910 -15.084  -1.874  12.729  74.124

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)       35.1270     0.5745  61.145  < 2e-16 ***
acousticness       0.7627     0.2246   3.396 0.000684 ***
danceability       5.5607     0.3603  15.433  < 2e-16 ***
energy            -6.6158     0.3871 -17.090  < 2e-16 ***
instrumentalness  -6.9978     0.1761 -39.740  < 2e-16 ***
liveness          -2.8486     0.3303  -8.625  < 2e-16 ***
loudness           0.7867     0.0148  53.160  < 2e-16 ***
speechiness       -5.9053     0.4480 -13.182  < 2e-16 ***
time_signature     0.7962     0.1045   7.623 2.49e-14 ***
valence           -7.0609     0.2377 -29.704  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 18.84 on 130653 degrees of freedom
Multiple R-squared:  0.08699,   Adjusted R-squared:  0.08693
F-statistic:  1383 on 9 and 130653 DF,  p-value: < 2.2e-16
```

Linear Models

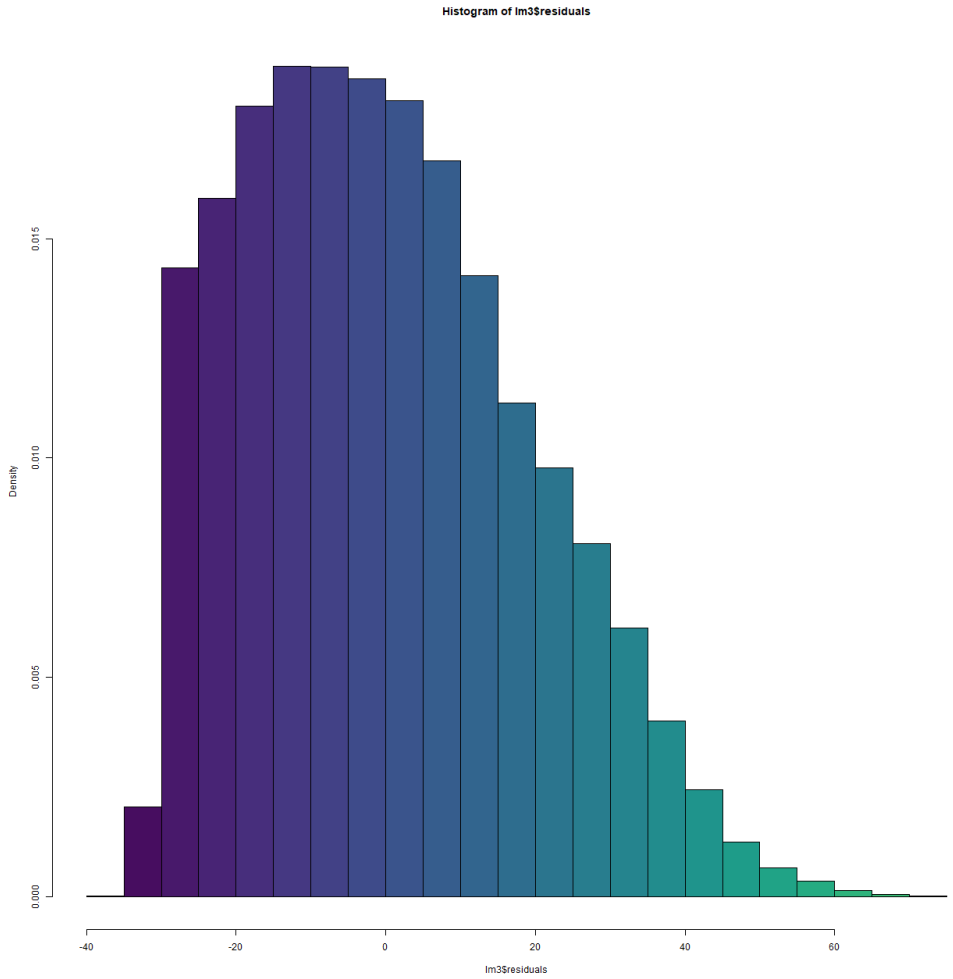
Reduced (Lowest correlation)

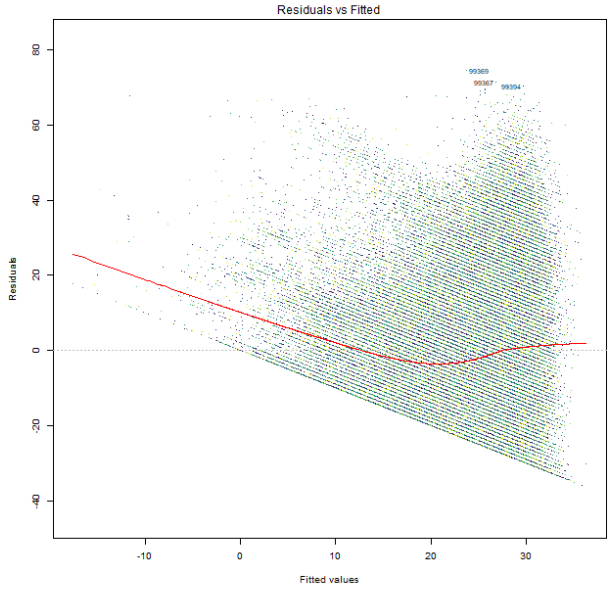
```
Call:
lm(formula = popularity ~ ., data = lm2)

Residuals:
    Min      1Q  Median      3Q     Max
-35.966 -15.101  -1.879  12.792  72.817

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)      36.163708   0.618575  58.463  < 2e-16 ***
acousticness      0.764781   0.225424   3.393 0.000692 ***
danceability      4.426231   0.352321  12.563  < 2e-16 ***
energy           -7.004398   0.387121 -18.094  < 2e-16 ***
instrumentalness -6.534236   0.172752 -37.824  < 2e-16 ***
liveness         -3.395957   0.328539 -10.337  < 2e-16 ***
loudness          0.820825   0.014654  56.013  < 2e-16 ***
mode             -0.105333   0.107456  -0.980 0.326969
tempo            -0.003800   0.001793  -2.120 0.034006 *
time_signature    0.796744   0.104625   7.615 2.65e-14 ***
valence          -6.968325   0.238427 -29.226  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 18.85 on 130652 degrees of freedom
Multiple R-squared:  0.08582,   Adjusted R-squared:  0.08575
F-statistic:  1226 on 10 and 130652 DF,  p-value: < 2.2e-16
```

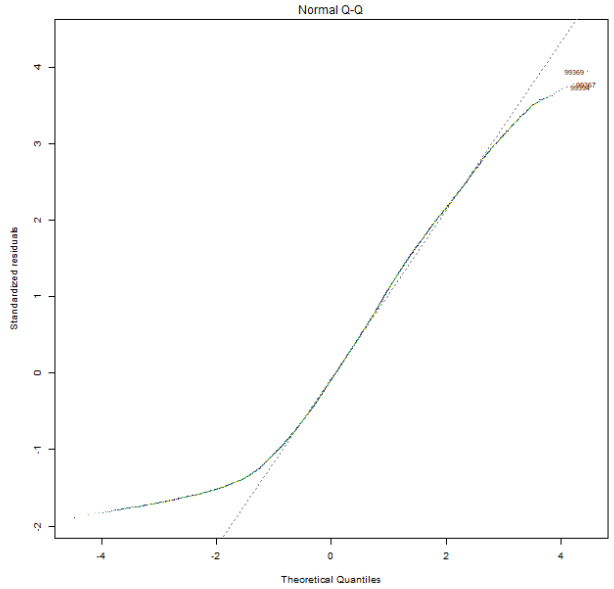
Reduced (Removed key)

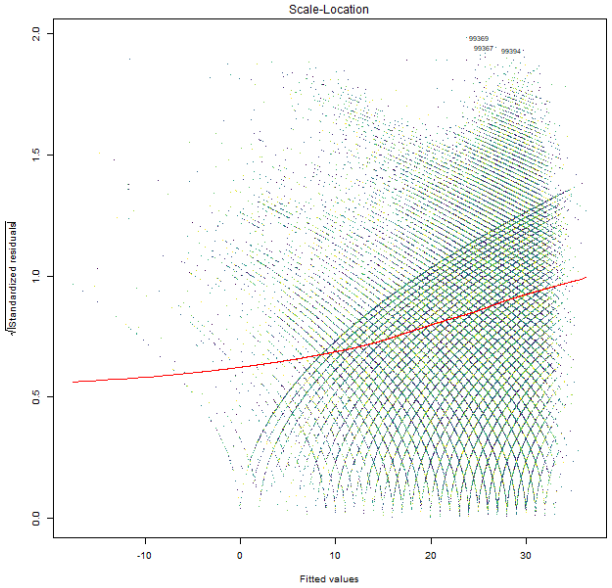
```
Call:
lm(formula = popularity ~ ., data = lm3)

Residuals:
    Min      1Q  Median      3Q     Max
-35.841 -15.095  -1.857  12.740  74.233

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)       3.579e+01  6.281e-01  56.982  < 2e-16 ***
acousticness      7.477e-01  2.253e-01   3.318 0.000906 ***
danceability      5.453e+00  3.622e-01  15.054  < 2e-16 ***
duration_ms      -8.414e-07  4.308e-07  -1.953 0.050803 .
energy           -6.615e+00  3.880e-01 -17.051  < 2e-16 ***
instrumentalness -7.002e+00  1.762e-01 -39.737  < 2e-16 ***
liveness         -2.878e+00  3.308e-01  -8.702  < 2e-16 ***
loudness          7.896e-01  1.485e-02  53.185  < 2e-16 ***
mode             -1.606e-01  1.075e-01  -1.495 0.134987
speechiness      -5.962e+00  4.499e-01 -13.252  < 2e-16 ***
tempo            -2.703e-03  1.793e-03  -1.507 0.131699
time_signature    8.088e-01  1.046e-01   7.729 1.09e-14 ***
valence          -7.072e+00  2.395e-01 -29.523  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 18.84 on 130650 degrees of freedom
Multiple R-squared:  0.08705,   Adjusted R-squared:  0.08697
F-statistic:  1038 on 12 and 130650 DF,  p-value: < 2.2e-16
```

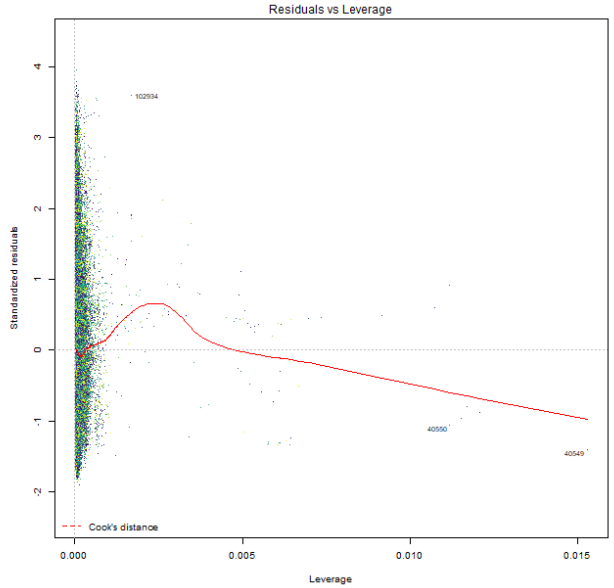
Residuals


Histogram of lm3$residuals
Density
0.000
0.005
0.010
0.015
-40
-20
0
20
40
60
lm3$residuals


Residuals vs Fitted
Residuals
Fitted values
99369
99367
99394


Normal Q-Q
Standardized residuals
Theoretical Quantiles
99369


Scale-Location
√|Standardized residuals|
Fitted values
99369
99367
99394


Residuals vs Leverage
Standardized residuals
Leverage
Cook's distance
102934
40550
40549

Conclusion

- No single variable can predict popularity
- Reduced Model is significant and can predict popularity
- Music is unique in that different people find different music popular
- Improvements to this model might be to add variables such as: times streamed, amount of time in top played, genre, production cost