Predicting the Popularity of a Song with Multiple Linear Regression
DATA 603 Fall 2020

Janny Tam, Joel Penner, Peter Baran

Introduction

The popularity of a song has been used as a metric to rank the success of artists early on in the music industry. In 1958, Billboard Hot 100 was introduced as a ranking system for the best-selling songs and is the industry standard for sales of music.

With the rise of streaming platforms, access to listening to a different variety of music has increased exponentially. One of the most popular streaming platforms for music today is Spotify. While record companies allow for their artist's songs to stream on Spotify for revenue, Spotify can break down and analyze each entry on their platform.

While assigning different metrics to the individual musical aspects of each song, Spotify also uses a 0-100 ranking for the popularity of each song. With this ranking system, are we able to find a pattern in what makes a song popular, or are we able to predict the popularity of a song based on its characteristics? Spotify and other music agencies may be able to use this model to find up and coming songs that would generate a better return on advertising knowing it would be a popular song.

Using the metrics that break down the characteristics of each song in our dataset, we hope to find the best model that is able to predict the popularity of a song based on its musical characteristics.To assist with our study, we have obtained the data for over 170,000 songs from Spotify through Kaggle.

Methodology

Descriptive Statistics

The dataset we will be using is ‘Spotify Dataset 1921-2020, 160k+ Tracks’ from Kaggle. We obtained permission to use this data by conforming to the Terms of Use of Kaggle. This dataset contains over 160,000 songs and many of their Audio Features. We will be using these Audio Features to predict the popularity of a song. Our dependent variable popularity ranges from 0 to 100. A higher popularity value means that the song is overall more popular. These Audio Features include 11 numerical variables and 3 categorical variables. The numerical variables are acousticness, danceability, energy, duration_ms, instrumentalness, valence, tempo, liveness, loudness, speechiness, and year. While the categorical variables are mode, explicit, and key.

Numerical Variables:

- Acousticness is a confidence measure from 0.0 to 1.0 to determine whether the track is acoustic. A measure of 1.0 means that there is high confidence that the track is acoustic.
- Danceability is a measure from 0.0 to 1.0 to determine how suitable a track is for dancing based on a combination of tempo, rhythm stability, beat strength, and overall regularity.

- Energy is a measure from 0.0 to 1.0 to determine the intensity and activity of a song. More energetic songs feel fast, loud, and noisy.
- Duration_ms is the duration of a song measured in milliseconds
- Instrumentalness is a measure from 0.0 to 1.0 that predicts whether a song contains vocals. The closer the instrumentalness value is to 1.0, the greater likelihood the song contains no vocal content.
- Valence is a measure from 0.0 to 1.0 determining the musical positiveness conveyed by a song. High valence sounds more positive while low valence sounds more negative.
- Tempo measures a song's beats per minute (BPM)
- Liveness is a measure from 0.0 to 1.0 determining the presence of an audience in the recording. Higher liveness values represent an increased probability that the track was performed live
- Loudness is measured in decibels (dB). Loudness is averaged across the entire song.
- Speechiness is a measure from 0.0 to 1.0 that detects the presence of spoken words in a song. Values before 0.33 most likely represent music.
- Year is the year the song was released

Categorical Variables:

- Mode indicates the modality of a song. Major is represented by 1 and minor is 0
- Explicit indicates if a song has explicit content. Explicit content is represented by 1 and no explicit content is represented by 0
- Key estimates the overall key of the track. The dataset uses integers map to pitches using standard Pitch Class notation. E.g. 0 = C, 1 = C♯/D♭, 2 = D, and so on. If there is no key detected, the value is -1.

Modelling

First, we will create a linear regression model using all 14 predictors and test for multicollinearity. After we have removed variables that have high multicollinearity, we will perform a stepwise regression to find the main effects of the model. If we find any main effects that have a p-value that is in the "gray area", we will perform partial F tests to assure that these main effects should be included in our model. With our main effects model, we can now find the interaction terms using individual t-tests. After we have found the significant interaction terms, we will once again run partial F tests on these interaction terms to determine if they should be concluded in the model. The next step will be to include any higher-order term that will improve the model. Once we have found a model that includes all the significant terms and has the highest R2adj and lowest RMSE, that will be considered our best fit model.

With our best fit model, we will now test our assumptions. First, we will test to see if the linearity assumption holds for our best fit model by looking at our Residuals vs Fitted plot. We will test the independence assumption by plotting the residuals against our Year variable. Next, we will test if our best fit model meets the equal variance assumption by using the Residuals vs Fitted plot, Scale-Location plot, and the Bresuch-Pagan test. The normality assumption will be checked by looking at the histogram of residuals and the Q-Q plot. Since the Shapiro-Wilk test is limited to a sample size of 3 to 5000, we decided to use the Anderson-Darling normality test.

We will test the multicollinearity assumption using the Variance Inflation Factor (VIF). Lastly, we will check if there are any outliers by using the Residuals vs Leverage plot, Cook's Distance, and Leverage Points.

Main Results of the Analysis

Multicollinearity

```
```{r}
imcdiag(spotify_fullmodel, method = 'VIF')
```
```

```
Call:
imcdiag(mod = spotify_fullmodel, method = "VIF")


 VIF Multicollinearity Diagnostics

                     VIF detection
valence           2.0520         0
acousticness      3.0608         0
danceability      1.9701         0
duration_ms       1.0774         0
energy            4.9624         0
factor(explicit)1 1.4139         0
instrumentalness  1.3596         0
factor(key)1      1.4958         0
factor(key)2      1.6686         0
factor(key)3      1.2900         0
factor(key)4      1.5008         0
factor(key)5      1.6046         0
factor(key)6      1.3516         0
factor(key)7      1.7247         0
factor(key)8      1.4062         0
factor(key)9      1.6415         0
factor(key)10     1.4625         0
factor(key)11     1.4466         0
liveness          1.1008         0
loudness          3.2056         0
factor(mode)1     1.0697         0
speechiness       1.5521         0
tempo             1.1106         0
year              2.0727         0

NOTE:  VIF Method Failed to detect multicollinearity


0 --> COLLINEARITY is not detected by the test

===================================
```

The VIF diagnostic shows that multicollinearity is not a significant problem for any of the variables, so we can start by building the first order model.

Stepwise for full model

```r
spotify_fullmodel = lm(popularity ~ valence + acousticness + danceability + duration_ms + energy +
factor(explicit) + instrumentalness + factor(key) + liveness + loudness + factor(mode) + speechiness + tempo +
year, data = spotify_data)
spotify_stepmodel = ols_step_both_p(spotify_fullmodel, pent = 0.05, prem = 0.1, details = F)
summary(spotify_stepmodel$model)
```

```
Call:
lm(formula = paste(response, "~", paste(preds, collapse = " + ")),
    data = l)

Residuals:
    Min      1Q  Median      3Q     Max
-64.165  -7.131  -1.376   5.654  69.371

Coefficients:
                   Estimate Std. Error  t value Pr(>|t|)
(Intercept)       -1.286e+03  2.886e+00 -445.740  < 2e-16 ***
acousticness      -4.099e+00  1.200e-01  -34.167  < 2e-16 ***
danceability       2.717e+00  2.061e-01   13.184  < 2e-16 ***
energy            -1.360e+00  1.689e-01   -8.055 8.00e-16 ***
factor(explicit)1  9.866e-01  1.114e-01    8.860  < 2e-16 ***
instrumentalness  -4.226e+00  9.218e-02  -45.840  < 2e-16 ***
factor(key)1      -1.738e-01  1.214e-01   -1.432 0.152130
factor(key)2      -1.092e-01  1.081e-01   -1.010 0.312520
factor(key)3      -1.795e-01  1.472e-01   -1.219 0.222747
factor(key)4       3.383e-01  1.214e-01    2.787 0.005327 **
factor(key)5       2.020e-01  1.126e-01    1.793 0.072943 .
factor(key)6       5.429e-02  1.383e-01    0.393 0.694631
factor(key)7      -1.158e-01  1.053e-01   -1.100 0.271242
factor(key)8       5.587e-02  1.280e-01    0.436 0.662495
factor(key)9       3.156e-01  1.106e-01    2.855 0.004310 **
factor(key)10     -1.028e-02  1.233e-01   -0.083 0.933547
factor(key)11     -4.944e-01  1.303e-01   -3.795 0.000148 ***
speechiness       -7.727e+00  1.961e-01  -39.412  < 2e-16 ***
tempo              1.806e-03  8.991e-04    2.008 0.044646 *
year               6.681e-01  1.454e-03  459.408  < 2e-16 ***
liveness          -2.996e+00  1.572e-01  -19.065  < 2e-16 ***
valence            6.444e-01  1.411e-01    4.566 4.98e-06 ***
factor(mode)1     -1.939e-01  5.953e-02   -3.257 0.001125 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 10.83 on 170630 degrees of freedom
Multiple R-squared:  0.7537,    Adjusted R-squared:  0.7536
F-statistic: 2.373e+04 on 22 and 170630 DF,  p-value: < 2.2e-16
```

The stepwise selection procedure uses t-tests for single β_i parameters:

$$H_0 : \beta_i = 0$$
$$H_A : \beta_i \neq 0$$

From our code you can see that pent = 0.05, meaning that variables with p-value less than 0.05 will enter the model. With prem = 0.1, variables with p-value more than 0.1 will be removed from the model. The stepwise regression procedure finds that the variables of duration_ms and loudness are not significant to be kept in the model. This leaves us with a first order model with the following regression equation:

$$\hat{Population} = \hat{\beta}_0 + \hat{\beta}_1 Acousticness + \hat{\beta}_2 Danceability + \hat{\beta}_3 Energy + \hat{\beta}_4 Explicit + \hat{\beta}_5 Instrumentalness + \hat{\beta}_6 Key + \hat{\beta}_7 Speechiness + \hat{\beta}_8 Tempo + \hat{\beta}_9 Year + \hat{\beta}_{10} Liveness + \hat{\beta}_{11} Valence + \hat{\beta}_{12} Mode$$

This first order model has an RMSE of 10.83, and an R^2_{adj} of 0.7536.

Interaction terms

```r
spotify_interactmodel = lm(popularity ~ (acousticness + danceability + energy + factor(explicit) + instrumentalness + factor(key) + speechiness + tempo + year + liveness + valence + factor(mode))^2, data = spotify_data)
summary(spotify_interactmodel)
```

```
Call:
lm(formula = popularity ~ (acousticness + danceability + energy +
    factor(explicit) + instrumentalness + factor(key) + speechiness +
    tempo + year + liveness + valence + factor(mode))^2, data = spotify_data)

Residuals:
   Min     1Q Median     3Q    Max
-66.37  -7.02  -1.37   5.58  68.92

Coefficients:
                                  Estimate Std. Error t value Pr(>|t|)
(Intercept)                     -1.361e+03  2.071e+01 -65.741  < 2e-16 ***
acousticness                    -3.123e+02  1.098e+01 -28.435  < 2e-16 ***
danceability                     3.293e+02  2.073e+01  15.886  < 2e-16 ***
energy                          -8.532e+00  1.691e+01  -0.504 0.613979
factor(explicit)1               -4.450e+02  1.569e+01 -28.360  < 2e-16 ***
instrumentalness                 1.500e+02  9.414e+00  15.939  < 2e-16 ***
factor(key)1                    -3.949e+01  1.294e+01  -3.053 0.002267 **
factor(key)2                    -9.997e+00  1.181e+01  -0.846 0.397331
factor(key)3                     5.496e+01  1.547e+01   3.553 0.000382 ***
factor(key)4                     2.228e+01  1.328e+01   1.677 0.093483 .
factor(key)5                    -2.607e+01  1.209e+01  -2.156 0.031083 *
factor(key)6                    -9.168e+00  1.500e+01  -0.611 0.540940
factor(key)7                     7.675e+00  1.140e+01   0.673 0.500904
factor(key)8                    -3.134e+01  1.381e+01  -2.269 0.023266 *
factor(key)9                     1.169e+01  1.220e+01   0.958 0.337950
factor(key)10                   -8.393e+00  1.334e+01  -0.629 0.529254
factor(key)11                    9.790e+00  1.444e+01   0.678 0.497730
speechiness                      4.174e+02  2.021e+01  20.657  < 2e-16 ***
tempo                            2.554e-01  9.342e-02   2.734 0.006256 **
year                             7.073e-01  1.043e-02  67.838  < 2e-16 ***
liveness                         5.330e+01  1.768e+01   3.015 0.002574 **
valence                          1.807e+01  1.411e+01   1.281 0.200367
factor(mode)1                    3.078e+01  6.486e+00   4.746 2.08e-06 ***
acousticness:danceability        8.673e+00  8.930e-01   9.712  < 2e-16 ***
acousticness:energy              5.755e+00  4.758e-01  12.097  < 2e-16 ***
acousticness:factor(explicit)1   1.460e+00  5.436e-01   2.685 0.007245 **
acousticness:instrumentalness    4.547e-01  4.321e-01   1.052 0.292729
acousticness:factor(key)1       -1.032e+00  5.438e-01  -1.898 0.057743 .
acousticness:factor(key)2       -8.615e-01  4.802e-01  -1.794 0.072839 .
acousticness:factor(key)3       -2.925e+00  7.186e-01  -4.071 4.68e-05 ***
acousticness:factor(key)4       -1.371e+00  5.443e-01  -2.519 0.011784 *
acousticness:factor(key)5       -1.408e+00  5.069e-01  -2.777 0.005486 **
acousticness:factor(key)6       -2.497e+00  6.186e-01  -4.036 5.43e-05 ***
acousticness:factor(key)7       -1.411e+00  4.674e-01  -3.019 0.002533 **
acousticness:factor(key)8       -5.170e-01  5.917e-01  -0.874 0.382256
acousticness:factor(key)9       -1.089e+00  4.888e-01  -2.228 0.025865 *
acousticness:factor(key)10      -2.004e+00  5.695e-01  -3.519 0.000433 ***
acousticness:factor(key)11      -1.826e+00  5.716e-01  -3.195 0.001399 **
acousticness:speechiness        -3.968e-01  7.885e-01  -0.503 0.614806
acousticness:tempo               1.513e-02  3.907e-03   3.872 0.000108 ***
acousticness:year                1.539e-01  5.560e-03  27.680  < 2e-16 ***
acousticness:liveness           -3.266e+00  6.723e-01  -4.858 1.19e-06 ***
acousticness:valence            -7.545e+00  5.903e-01 -12.781  < 2e-16 ***
acousticness:factor(mode)1       3.011e-01  2.702e-01   1.114 0.265073
danceability:energy              2.487e-01  1.237e+00   0.201 0.840672
danceability:factor(explicit)1   5.811e+00  8.257e-01   7.037 1.97e-12 ***
danceability:instrumentalness   -1.181e+01  6.550e-01 -18.036  < 2e-16 ***
danceability:factor(key)1       -2.099e+00  9.079e-01  -2.312 0.020770 *
danceability:factor(key)2        2.105e-01  8.393e-01   0.251 0.801984
danceability:factor(key)3        1.621e+00  1.172e+00   1.383 0.166737
danceability:factor(key)4        1.442e+00  9.436e-01   1.528 0.126463
danceability:factor(key)5        5.119e-01  8.871e-01   0.577 0.563923
danceability:factor(key)6       -3.816e-01  1.065e+00  -0.358 0.720188
danceability:factor(key)7       -1.245e+00  8.208e-01  -1.517 0.129169
danceability:factor(key)8        8.990e-01  9.994e-01   0.900 0.368354
danceability:factor(key)9       -3.016e-01  8.764e-01  -0.344 0.730740
danceability:factor(key)10       3.474e-01  9.623e-01   0.361 0.718057
danceability:factor(key)11      -1.899e+00  1.004e+00  -1.891 0.058617 .
danceability:speechiness         3.854e+00  2.180e+00   1.768 0.077010 .
danceability:tempo              -1.833e-02  6.882e-03  -2.664 0.007724 **
danceability:year               -1.646e-01  1.035e-02 -15.904  < 2e-16 ***
danceability:liveness            1.420e+00  1.237e+00   1.148 0.250773
danceability:valence            -1.526e+00  7.061e-01  -2.161 0.030664 *
danceability:factor(mode)1       4.305e-01  4.589e-01   0.938 0.348169
```

energy:factor(explicit)1	-3.081e+00	7.237e-01	-4.257	2.07e-05	***
energy:instrumentalness	-5.335e+00	5.936e-01	-8.987	< 2e-16	***
energy:factor(key)1	-2.073e+00	7.591e-01	-2.730	0.006326	**
energy:factor(key)2	-7.146e-01	6.778e-01	-1.054	0.291742	
energy:factor(key)3	-9.012e-01	9.729e-01	-0.926	0.354282	
energy:factor(key)4	-1.008e+00	7.678e-01	-1.313	0.189180	
energy:factor(key)5	-2.044e+00	7.222e-01	-2.830	0.004652	**
energy:factor(key)6	-2.137e+00	8.749e-01	-2.443	0.014560	*
energy:factor(key)7	-2.445e+00	6.664e-01	-3.669	0.000243	***
energy:factor(key)8	-9.481e-01	8.198e-01	-1.157	0.247472	
energy:factor(key)9	-1.357e+00	6.984e-01	-1.943	0.051998	.
energy:factor(key)10	-2.827e+00	7.937e-01	-3.562	0.000369	***
energy:factor(key)11	-2.261e+00	8.057e-01	-2.806	0.005010	**
energy:speechiness	1.674e+01	1.328e+00	12.608	< 2e-16	***
energy:tempo	-1.082e-02	5.393e-03	-2.006	0.044836	*
energy:year	3.525e-03	8.502e-03	0.415	0.678437	
energy:liveness	-2.545e+00	8.748e-01	-2.909	0.003624	**
energy:valence	2.304e+00	7.585e-01	3.037	0.002390	**
energy:factor(mode)1	1.298e+00	3.780e-01	3.433	0.000598	***
factor(explicit)1:instrumentalness	6.281e+00	8.732e-01	7.194	6.34e-13	***
factor(explicit)1:factor(key)1	-1.070e+00	4.680e-01	-2.286	0.022274	*
factor(explicit)1:factor(key)2	-1.668e+00	5.153e-01	-3.238	0.001204	**
factor(explicit)1:factor(key)3	-9.670e-01	8.087e-01	-1.196	0.231783	
factor(explicit)1:factor(key)4	-9.814e-01	5.591e-01	-1.755	0.079218	.
factor(explicit)1:factor(key)5	-8.364e-01	5.569e-01	-1.502	0.133126	
factor(explicit)1:factor(key)6	-1.141e+00	5.319e-01	-2.145	0.031981	*
factor(explicit)1:factor(key)7	1.002e-01	4.901e-01	0.204	0.838086	
factor(explicit)1:factor(key)8	-7.863e-01	5.477e-01	-1.436	0.151137	
factor(explicit)1:factor(key)9	-8.627e-01	5.189e-01	-1.662	0.096431	.
factor(explicit)1:factor(key)10	-7.577e-01	5.560e-01	-1.363	0.172922	
factor(explicit)1:factor(key)11	-2.205e-01	5.095e-01	-0.433	0.665103	
factor(explicit)1:speechiness	8.131e+00	7.440e-01	10.929	< 2e-16	***
factor(explicit)1:tempo	1.974e-02	3.682e-03	5.363	8.22e-08	***
factor(explicit)1:year	2.208e-01	7.866e-03	28.074	< 2e-16	***
factor(explicit)1:liveness	1.067e+00	6.504e-01	1.641	0.100852	
factor(explicit)1:valence	-1.642e+00	5.555e-01	-2.956	0.003119	**
factor(explicit)1:factor(mode)1	7.218e-01	2.445e-01	2.952	0.003155	**
instrumentalness:factor(key)1	5.010e-02	4.225e-01	0.119	0.905618	
instrumentalness:factor(key)2	7.807e-01	3.719e-01	2.099	0.035829	*
instrumentalness:factor(key)3	-6.084e-01	4.658e-01	-1.306	0.191510	
instrumentalness:factor(key)4	1.183e-01	4.236e-01	0.279	0.780016	
instrumentalness:factor(key)5	6.460e-01	3.742e-01	1.726	0.084276	.
instrumentalness:factor(key)6	8.299e-02	4.994e-01	0.166	0.868010	
instrumentalness:factor(key)7	1.232e-01	3.613e-01	0.341	0.733205	
instrumentalness:factor(key)8	4.484e-01	4.304e-01	1.042	0.297508	
instrumentalness:factor(key)9	5.552e-01	3.912e-01	1.419	0.155838	
instrumentalness:factor(key)10	-3.827e-01	4.146e-01	-0.923	0.356078	
instrumentalness:factor(key)11	-3.954e-01	4.752e-01	-0.832	0.405290	
instrumentalness:speechiness	6.170e-01	1.395e+00	0.442	0.658334	
instrumentalness:tempo	-1.670e-02	3.038e-03	-5.498	3.85e-08	***
instrumentalness:year	-7.528e-02	4.713e-03	-15.973	< 2e-16	***
instrumentalness:liveness	1.879e+00	5.827e-01	3.225	0.001261	**
instrumentalness:valence	6.065e+00	4.577e-01	13.250	< 2e-16	***
instrumentalness:factor(mode)1	8.039e-02	2.009e-01	0.400	0.689060	
factor(key)1:speechiness	1.538e+00	8.222e-01	1.871	0.061387	.
factor(key)2:speechiness	6.563e-01	9.058e-01	0.725	0.468727	
factor(key)3:speechiness	2.464e+00	1.193e+00	2.065	0.038908	*
factor(key)4:speechiness	2.472e-01	1.020e+00	0.242	0.808482	
factor(key)5:speechiness	1.170e-01	9.251e-01	0.127	0.899319	
factor(key)6:speechiness	2.002e+00	9.179e-01	2.181	0.029213	*
factor(key)7:speechiness	7.375e-01	8.431e-01	0.875	0.381705	
factor(key)8:speechiness	1.085e+00	9.890e-01	1.097	0.272641	
factor(key)9:speechiness	1.699e+00	8.581e-01	1.980	0.047701	*
factor(key)10:speechiness	3.124e+00	9.143e-01	3.416	0.000635	***
factor(key)11:speechiness	1.005e+00	8.873e-01	1.132	0.257465	
factor(key)1:tempo	2.934e-03	4.044e-03	0.726	0.468124	
factor(key)2:tempo	2.671e-04	3.680e-03	0.073	0.942137	
factor(key)3:tempo	4.070e-03	4.846e-03	0.840	0.400913	
factor(key)4:tempo	7.714e-03	4.121e-03	1.872	0.061231	.
factor(key)5:tempo	5.545e-03	3.767e-03	1.472	0.141019	
factor(key)6:tempo	-2.118e-03	4.660e-03	-0.455	0.649435	
factor(key)7:tempo	2.478e-03	3.570e-03	0.694	0.487638	
factor(key)8:tempo	4.677e-03	4.256e-03	1.099	0.271762	
factor(key)9:tempo	9.402e-03	3.768e-03	2.495	0.012598	*
factor(key)10:tempo	4.012e-03	4.122e-03	0.973	0.330389	
factor(key)11:tempo	3.450e-03	4.441e-03	0.777	0.437208	
factor(key)1:year	2.085e-02	6.510e-03	3.203	0.001358	**
factor(key)2:year	5.361e-03	5.954e-03	0.900	0.367924	
factor(key)3:year	-2.743e-02	7.760e-03	-3.535	0.000408	***
factor(key)4:year	-1.129e-02	6.700e-03	-1.685	0.092004	.
factor(key)5:year	1.357e-02	6.092e-03	2.228	0.025868	*
factor(key)6:year	6.022e-03	7.560e-03	0.797	0.425683	
factor(key)7:year	-3.015e-03	5.752e-03	-0.524	0.600186	
factor(key)8:year	1.570e-02	6.958e-03	2.256	0.024051	*
factor(key)9:year	-5.775e-03	6.160e-03	-0.938	0.348493	
factor(key)10:year	5.106e-03	6.716e-03	0.760	0.447053	

```
factor(key)11:year                -4.371e-03  7.294e-03  -0.599 0.549019
factor(key)1:liveness              1.965e-01  7.195e-01   0.273 0.784811
factor(key)2:liveness             -5.473e-01  6.266e-01  -0.873 0.382471
factor(key)3:liveness              1.351e+00  9.414e-01   1.435 0.151179
factor(key)4:liveness              2.788e-01  7.028e-01   0.397 0.691612
factor(key)5:liveness              4.111e-01  6.695e-01   0.614 0.539241
factor(key)6:liveness             -9.202e-01  8.272e-01  -1.112 0.265952
factor(key)7:liveness              4.637e-01  6.024e-01   0.770 0.441499
factor(key)8:liveness             -5.867e-01  7.902e-01  -0.742 0.457835
factor(key)9:liveness             -7.376e-01  6.435e-01  -1.146 0.251735
factor(key)10:liveness            -1.100e-01  7.401e-01  -0.149 0.881808
factor(key)11:liveness             9.740e-01  7.479e-01   1.302 0.192800
factor(key)1:valence               1.112e+00  6.319e-01   1.759 0.078541 .
factor(key)2:valence              -6.549e-01  5.744e-01  -1.140 0.254277
factor(key)3:valence              -1.291e+00  8.066e-01  -1.600 0.109513
factor(key)4:valence              -9.970e-01  6.413e-01  -1.555 0.120017
factor(key)5:valence              -1.776e-01  6.104e-01  -0.291 0.771049
factor(key)6:valence              -1.655e-01  7.344e-01  -0.225 0.821722
factor(key)7:valence               1.402e-01  5.600e-01   0.250 0.802255
factor(key)8:valence              -1.029e+00  6.757e-01  -1.522 0.127935
factor(key)9:valence              -2.862e-01  5.935e-01  -0.482 0.629647
factor(key)10:valence             -8.965e-01  6.572e-01  -1.364 0.172519
factor(key)11:valence              1.339e+00  6.844e-01   1.957 0.050341 .
factor(key)1:factor(mode)1         3.591e-02  2.865e-01   0.125 0.900254
factor(key)2:factor(mode)1         3.631e-01  2.656e-01   1.367 0.171575
factor(key)3:factor(mode)1         4.580e-01  3.674e-01   1.246 0.212599
factor(key)4:factor(mode)1         4.790e-01  2.678e-01   1.789 0.073661 .
factor(key)5:factor(mode)1         2.791e-01  2.575e-01   1.084 0.278511
factor(key)6:factor(mode)1         5.617e-01  3.002e-01   1.871 0.061361 .
factor(key)7:factor(mode)1         2.449e-01  2.628e-01   0.932 0.351418
factor(key)8:factor(mode)1         1.036e+00  3.257e-01   3.182 0.001463 **
factor(key)9:factor(mode)1         3.148e-01  2.535e-01   1.242 0.214306
factor(key)10:factor(mode)1        6.459e-01  2.891e-01   2.234 0.025489 *
factor(key)11:factor(mode)1        4.616e-01  2.833e-01   1.630 0.103188
speechiness:tempo                  1.778e-03  6.367e-03   0.279 0.780095
speechiness:year                  -2.231e-01  1.041e-02 -21.424  < 2e-16 ***
speechiness:liveness               1.134e+00  1.020e+00   1.113 0.265883
speechiness:valence                3.345e-01  1.044e+00   0.320 0.748624
speechiness:factor(mode)1         -7.430e-01  4.435e-01  -1.675 0.093892 .
tempo:year                        -1.311e-04  4.695e-05  -2.792 0.005246 **
tempo:liveness                    -1.006e-02  5.255e-03  -1.915 0.055539 .
tempo:valence                      1.589e-02  4.524e-03   3.512 0.000445 ***
tempo:factor(mode)1                3.993e-03  2.008e-03   1.988 0.046811 *
year:liveness                     -2.798e-02  8.920e-03  -3.137 0.001708 **
year:valence                      -9.198e-03  7.126e-03  -1.291 0.196823
year:factor(mode)1                -1.650e-02  3.268e-03  -5.047 4.48e-07 ***
liveness:valence                   2.965e+00  8.258e-01   3.591 0.000329 ***
liveness:factor(mode)1             9.063e-01  3.563e-01   2.544 0.010965 *
valence:factor(mode)1             -8.786e-01  3.151e-01  -2.789 0.005293 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 10.62 on 170454 degrees of freedom
Multiple R-squared:  0.7634,    Adjusted R-squared:  0.7631
F-statistic:  2778 on 198 and 170454 DF,  p-value: < 2.2e-16
```

The first take on an interaction model has a staggering 176 interaction terms. We will use individual t-tests to determine if an interaction term should be included in our model:

$$H_0 : \beta_i = 0$$
$$H_A : \beta_i \neq 0$$

With a level of significance of 0.05, we can see that there are a lot of terms that are greater than 0.05. Therefore we fail to reject the null hypothesis and state that those interaction terms are not significant to be kept in the model. Any interaction terms that are in the “gray area”, we will be doing partial F tests to verify that they are significant in the model.

```
Model 1: popularity ~ acousticness + danceability + energy + factor(explicit) +
    instrumentalness + factor(key) + speechiness + tempo + year +
    liveness + valence + factor(mode) + acousticness * danceability +
    acousticness * energy + acousticness * factor(explicit) +
    acousticness * factor(key) + acousticness * tempo + acousticness *
    year + acousticness * liveness + acousticness * valence +
    danceability * factor(explicit) + danceability * instrumentalness +
    danceability * speechiness + danceability * tempo + danceability *
    year + danceability * valence + energy * factor(explicit) +
    energy * instrumentalness + energy * factor(key) + energy *
    speechiness + energy * tempo + energy * liveness + energy *
    valence + energy * factor(mode) + factor(explicit) * instrumentalness +
    factor(explicit) * factor(key) + factor(explicit) * speechiness +
    factor(explicit) * tempo + factor(explicit) * year + factor(explicit) *
    valence + factor(explicit) * factor(mode) + instrumentalness *
    factor(key) + instrumentalness * tempo + instrumentalness *
    year + instrumentalness * liveness + instrumentalness * valence +
    factor(key) * speechiness + factor(key) * year + factor(key) *
    valence + factor(key) * factor(mode) + speechiness * year +
    speechiness * factor(mode) + tempo * year + tempo * liveness +
    tempo * valence + tempo * factor(mode) + year * liveness +
    year * factor(mode) + liveness * valence + liveness * factor(mode) +
    valence * factor(mode)
Model 2: popularity ~ acousticness + danceability + energy + factor(explicit) +
    instrumentalness + factor(key) + speechiness + tempo + year +
    liveness + valence + factor(mode) + acousticness * danceability +
    acousticness * energy + acousticness * factor(explicit) +
    acousticness * factor(key) + acousticness * tempo + acousticness *
    year + acousticness * liveness + acousticness * valence +
    danceability * factor(explicit) + danceability * instrumentalness +
    danceability * factor(key) + danceability * speechiness +
    danceability * tempo + danceability * year + danceability *
    valence + energy * factor(explicit) + energy * instrumentalness +
    energy * factor(key) + energy * speechiness + energy * tempo +
    energy * liveness + energy * valence + energy * factor(mode) +
    factor(explicit) * instrumentalness + factor(explicit) *
    factor(key) + factor(explicit) * speechiness + factor(explicit) *
    tempo + factor(explicit) * year + factor(explicit) * valence +
    factor(explicit) * factor(mode) + instrumentalness * factor(key) +
    instrumentalness * tempo + instrumentalness * year + instrumentalness *
    liveness + instrumentalness * valence + factor(key) * speechiness +
    factor(key) * year + factor(key) * valence + factor(key) *
    factor(mode) + speechiness * year + speechiness * factor(mode) +
    tempo * year + tempo * liveness + tempo * valence + tempo *
    factor(mode) + year * liveness + year * factor(mode) + liveness *
    valence + liveness * factor(mode) + valence * factor(mode)
  Res.Df      RSS Df Sum of Sq      F   Pr(>F)
1 170501 19242757
2 170490 19239517 11    3240.6 2.6106 0.002513 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Since we have several interaction terms that are in the “gray area”, we will be showing a partial F test for only one of these variables. From the summary, danceability*key were only significant when the level of key was 1 and 11. Therefore we perform a partial F test to confirm if danceability*key should be kept in the model. The partial F test has hypothesis:

$$H_0 : \beta_i = 0 \textit{ in the model}$$
$$H_A : \beta_i \neq 0 \textit{ in the model}$$

The figure above shows the partial F test for the interaction between danceability and key. With a level of significance of 0.05 and a p-value of 0.002513, we reject the null hypothesis and conclude that danceability*key is significant in the model. This test was repeated as needed until we arrived at the final interaction model shown below.

```{r}
spotify_interactionmodel = lm(popularity ~ acousticness + I(acousticness^2) + danceability + energy +
factor(explicit) + instrumentalness + factor(key) + speechiness + tempo + year + liveness + valence +
factor(mode) + acousticness*danceability + acousticness*energy + acousticness*factor(explicit) +
acousticness*factor(key) + acousticness*tempo + acousticness*year + acousticness*liveness  +
acousticness*valence + danceability*factor(explicit) + danceability*instrumentalness +
danceability*factor(key) + danceability*tempo + danceability*year + danceability*valence +
energy*factor(explicit) + energy*instrumentalness + energy*factor(key) + energy*speechiness + energy*tempo +
energy*liveness + energy*valence + energy*factor(mode) + factor(explicit)*instrumentalness +
factor(explicit)*factor(key) + factor(explicit)*speechiness + factor(explicit)*tempo + factor(explicit)*year +
factor(explicit)*valence  + factor(explicit)*factor(mode) + instrumentalness*factor(key) +
instrumentalness*tempo + instrumentalness*year + instrumentalness*liveness + instrumentalness*valence +
factor(key)*speechiness + factor(key)*year + factor(key)*valence + speechiness*year + tempo*year +
tempo*liveness  + tempo*valence + year*liveness + year*factor(mode) + liveness*valence + liveness*factor(mode)
+ valence*factor(mode), data = spotify_data)
summary(spotify_interactionmodel)
```

```
Call:
lm(formula = popularity ~ acousticness + I(acousticness^2) +
    danceability + energy + factor(explicit) + instrumentalness +
    factor(key) + speechiness + tempo + year + liveness + valence +
    factor(mode) + acousticness * danceability + acousticness *
    energy + acousticness * factor(explicit) + acousticness *
    factor(key) + acousticness * tempo + acousticness * year +
    acousticness * liveness + acousticness * valence + danceability *
    factor(explicit) + danceability * instrumentalness + danceability *
    factor(key) + danceability * tempo + danceability * year +
    danceability * valence + energy * factor(explicit) + energy *
    instrumentalness + energy * factor(key) + energy * speechiness +
    energy * tempo + energy * liveness + energy * valence + energy *
    factor(mode) + factor(explicit) * instrumentalness + factor(explicit) *
    factor(key) + factor(explicit) * speechiness + factor(explicit) *
    tempo + factor(explicit) * year + factor(explicit) * valence +
    factor(explicit) * factor(mode) + instrumentalness * factor(key) +
    instrumentalness * tempo + instrumentalness * year + instrumentalness *
    liveness + instrumentalness * valence + factor(key) * speechiness +
    factor(key) * year + factor(key) * valence + speechiness *
    year + tempo * year + tempo * liveness + tempo * valence +
    year * liveness + year * factor(mode) + liveness * valence +
    liveness * factor(mode) + valence * factor(mode), data = spotify_data)

Residuals:
    Min      1Q  Median      3Q     Max
-65.764  -7.009  -1.344   5.563  67.647
```

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	-1.435e+03	1.841e+01	-77.938	< 2e-16	***
acousticness	-2.264e+02	9.025e+00	-25.090	< 2e-16	***
I(acousticness^2)	-1.115e+01	4.618e-01	-24.141	< 2e-16	***
danceability	3.941e+02	1.687e+01	23.368	< 2e-16	***
energy	2.166e+00	9.231e-01	2.347	0.018939	*
factor(explicit)1	-4.313e+02	1.557e+01	-27.704	< 2e-16	***
instrumentalness	1.473e+02	8.447e+00	17.439	< 2e-16	***
factor(key)1	-3.723e+01	1.285e+01	-2.898	0.003754	**
factor(key)2	-6.257e+00	1.175e+01	-0.533	0.594232	
factor(key)3	5.084e+01	1.540e+01	3.301	0.000964	***
factor(key)4	2.429e+01	1.317e+01	1.844	0.065221	.
factor(key)5	-2.727e+01	1.201e+01	-2.271	0.023169	*
factor(key)6	-6.000e+00	1.490e+01	-0.403	0.687137	
factor(key)7	8.495e+00	1.133e+01	0.750	0.453312	
factor(key)8	-3.129e+01	1.371e+01	-2.283	0.022458	*
factor(key)9	1.357e+01	1.212e+01	1.120	0.262668	
factor(key)10	-9.349e+00	1.324e+01	-0.706	0.480255	
factor(key)11	1.672e+01	1.430e+01	1.169	0.242282	
speechiness	4.070e+02	1.979e+01	20.569	< 2e-16	***
tempo	3.662e-01	8.673e-02	4.222	2.43e-05	***
year	7.414e-01	9.260e-03	80.065	< 2e-16	***
liveness	6.009e+01	1.651e+01	3.640	0.000273	***
valence	-5.449e-01	9.540e-01	-0.571	0.567854	
factor(mode)1	3.193e+01	5.605e+00	5.698	1.22e-08	***
acousticness:danceability	5.983e+00	6.675e-01	8.963	< 2e-16	***
acousticness:energy	-8.806e-01	5.149e-01	-1.710	0.087189	.
acousticness:factor(explicit)1	-5.705e-01	4.737e-01	-1.204	0.228482	
acousticness:factor(key)1	-1.136e+00	5.388e-01	-2.109	0.034927	*
acousticness:factor(key)2	-9.841e-01	4.766e-01	-2.065	0.038960	*
acousticness:factor(key)3	-1.832e+00	7.073e-01	-2.590	0.009609	**
acousticness:factor(key)4	-1.410e+00	5.354e-01	-2.633	0.008461	**
acousticness:factor(key)5	-1.010e+00	5.021e-01	-2.012	0.044187	*
acousticness:factor(key)6	-2.590e+00	6.093e-01	-4.251	2.13e-05	***
acousticness:factor(key)7	-1.346e+00	4.640e-01	-2.901	0.003716	**
acousticness:factor(key)8	-1.746e-01	5.883e-01	-0.297	0.766624	
acousticness:factor(key)9	-1.424e+00	4.835e-01	-2.945	0.003233	**
acousticness:factor(key)10	-1.592e+00	5.606e-01	-2.839	0.004521	**
acousticness:factor(key)11	-2.086e+00	5.559e-01	-3.752	0.000176	***
acousticness:tempo	8.276e-03	3.834e-03	2.158	0.030907	*
acousticness:year	1.184e-01	4.465e-03	26.514	< 2e-16	***
acousticness:liveness	-2.500e+00	6.421e-01	-3.893	9.91e-05	***
acousticness:valence	-6.802e+00	5.139e-01	-13.235	< 2e-16	***
danceability:factor(explicit)1	5.837e+00	7.597e-01	7.683	1.56e-14	***
danceability:instrumentalness	-1.173e+01	6.331e-01	-18.523	< 2e-16	***
danceability:factor(key)1	-2.217e+00	8.907e-01	-2.489	0.012805	*
danceability:factor(key)2	1.665e-01	8.188e-01	0.203	0.838858	
danceability:factor(key)3	1.349e+00	1.148e+00	1.175	0.240006	
danceability:factor(key)4	9.441e-01	9.115e-01	1.036	0.300291	
danceability:factor(key)5	3.403e-01	8.635e-01	0.394	0.693476	
danceability:factor(key)6	-2.490e-01	1.037e+00	-0.240	0.810232	
danceability:factor(key)7	-1.454e+00	7.985e-01	-1.821	0.068683	.
danceability:factor(key)8	8.349e-01	9.774e-01	0.854	0.392991	
danceability:factor(key)9	-8.546e-01	8.484e-01	-1.007	0.313737	
danceability:factor(key)10	2.910e-01	9.361e-01	0.311	0.755877	
danceability:factor(key)11	-2.327e+00	9.672e-01	-2.406	0.016130	*
danceability:tempo	-1.857e-02	6.568e-03	-2.827	0.004696	**
danceability:year	-1.962e-01	8.403e-03	-23.355	< 2e-16	***
danceability:valence	-1.735e+00	6.524e-01	-2.659	0.007835	**

```
energy:factor(explicit)1           -5.087e+00  6.701e-01  -7.591 3.19e-14 ***
energy:instrumentalness            -5.702e+00  4.410e-01 -12.928  < 2e-16 ***
energy:factor(key)1                -2.087e+00  7.384e-01  -2.827 0.004701 **
energy:factor(key)2                -8.396e-01  6.632e-01  -1.266 0.205554
energy:factor(key)3                 3.536e-01  9.458e-01   0.374 0.708493
energy:factor(key)4                -8.874e-01  7.483e-01  -1.186 0.235648
energy:factor(key)5                -1.496e+00  7.039e-01  -2.126 0.033519 *
energy:factor(key)6                -2.419e+00  8.481e-01  -2.853 0.004336 **
energy:factor(key)7                -2.230e+00  6.531e-01  -3.415 0.000638 ***
energy:factor(key)8                -7.280e-01  8.023e-01  -0.907 0.364200
energy:factor(key)9                -1.548e+00  6.832e-01  -2.266 0.023464 *
energy:factor(key)10               -2.551e+00  7.735e-01  -3.298 0.000974 ***
energy:factor(key)11               -2.347e+00  7.828e-01  -2.998 0.002718 **
energy:speechiness                  1.830e+01  1.166e+00  15.702  < 2e-16 ***
energy:tempo                       -1.575e-02  5.307e-03  -2.967 0.003008 **
energy:liveness                    -1.922e+00  8.309e-01  -2.313 0.020747 *
energy:valence                      1.642e+00  6.079e-01   2.700 0.006932 **
energy:factor(mode)1                1.209e+00  2.958e-01   4.089 4.34e-05 ***
factor(explicit)1:instrumentalness  6.085e+00  8.679e-01   7.011 2.37e-12 ***
factor(explicit)1:factor(key)1     -1.165e+00  4.664e-01  -2.498 0.012490 *
factor(explicit)1:factor(key)2     -1.686e+00  5.141e-01  -3.279 0.001042 **
factor(explicit)1:factor(key)3     -1.204e+00  8.022e-01  -1.501 0.133260
factor(explicit)1:factor(key)4     -1.032e+00  5.559e-01  -1.856 0.063425 .
factor(explicit)1:factor(key)5     -7.922e-01  5.534e-01  -1.432 0.152273
factor(explicit)1:factor(key)6     -1.226e+00  5.295e-01  -2.316 0.020569 *
factor(explicit)1:factor(key)7      8.550e-02  4.890e-01   0.175 0.861192
factor(explicit)1:factor(key)8     -9.035e-01  5.459e-01  -1.655 0.097888 .
factor(explicit)1:factor(key)9     -8.536e-01  5.169e-01  -1.651 0.098681 .
factor(explicit)1:factor(key)10    -8.715e-01  5.503e-01  -1.584 0.113268
factor(explicit)1:factor(key)11    -3.039e-01  5.071e-01  -0.599 0.548927
factor(explicit)1:speechiness       8.147e+00  7.083e-01  11.502  < 2e-16 ***
factor(explicit)1:tempo             1.890e-02  3.339e-03   5.661 1.50e-08 ***
factor(explicit)1:year              2.151e-01  7.798e-03  27.578  < 2e-16 ***
factor(explicit)1:valence          -1.544e+00  5.132e-01  -3.008 0.002630 **
factor(explicit)1:factor(mode)1     6.403e-01  2.190e-01   2.924 0.003456 **
instrumentalness:factor(key)1      -3.322e-02  4.203e-01  -0.079 0.937005
instrumentalness:factor(key)2       6.631e-01  3.698e-01   1.793 0.072986 .
instrumentalness:factor(key)3      -7.516e-01  4.627e-01  -1.624 0.104310
instrumentalness:factor(key)4      -5.097e-02  4.176e-01  -0.122 0.902873
instrumentalness:factor(key)5       5.364e-01  3.713e-01   1.444 0.148632
instrumentalness:factor(key)6       1.585e-03  4.962e-01   0.003 0.997452
instrumentalness:factor(key)7       5.067e-02  3.590e-01   0.141 0.887768
instrumentalness:factor(key)8       3.011e-01  4.279e-01   0.704 0.481668
instrumentalness:factor(key)9       4.621e-01  3.878e-01   1.192 0.233349
instrumentalness:factor(key)10     -5.291e-01  4.101e-01  -1.290 0.196956
instrumentalness:factor(key)11     -5.857e-01  4.694e-01  -1.248 0.212143
instrumentalness:tempo             -1.622e-02  3.007e-03  -5.394 6.90e-08 ***
instrumentalness:year              -7.350e-02  4.309e-03 -17.055  < 2e-16 ***
instrumentalness:liveness           1.600e+00  5.664e-01   2.825 0.004731 **
instrumentalness:valence            6.561e+00  4.432e-01  14.804  < 2e-16 ***
factor(key)1:speechiness            1.704e+00  8.106e-01   2.102 0.035559 *
factor(key)2:speechiness            6.831e-01  8.928e-01   0.765 0.444173
factor(key)3:speechiness            3.888e+00  1.140e+00   3.411 0.000649 ***
factor(key)4:speechiness            4.140e-01  1.002e+00   0.413 0.679580
factor(key)5:speechiness            9.027e-01  8.993e-01   1.004 0.315491
factor(key)6:speechiness            1.869e+00  8.992e-01   2.078 0.037672 *
factor(key)7:speechiness            8.708e-01  8.302e-01   1.049 0.294236
factor(key)8:speechiness            1.405e+00  9.736e-01   1.443 0.148990
factor(key)9:speechiness            1.886e+00  8.375e-01   2.252 0.024304 *
factor(key)10:speechiness           3.723e+00  8.727e-01   4.266 1.99e-05 ***
factor(key)11:speechiness           1.277e+00  8.516e-01   1.500 0.133618
```

```
factor(key)1:year                 1.997e-02  6.468e-03   3.087 0.002024 **
factor(key)2:year                 3.641e-03  5.916e-03   0.615 0.538280
factor(key)3:year                -2.535e-02  7.716e-03  -3.285 0.001020 **
factor(key)4:year                -1.161e-02  6.635e-03  -1.749 0.080224 .
factor(key)5:year                 1.443e-02  6.045e-03   2.387 0.017005 *
factor(key)6:year                 4.476e-03  7.508e-03   0.596 0.551055
factor(key)7:year                -3.176e-03  5.709e-03  -0.556 0.577973
factor(key)8:year                 1.615e-02  6.901e-03   2.340 0.019278 *
factor(key)9:year                -5.932e-03  6.111e-03  -0.971 0.331695
factor(key)10:year                5.843e-03  6.662e-03   0.877 0.380423
factor(key)11:year               -7.237e-03  7.219e-03  -1.002 0.316113
factor(key)1:valence              1.316e+00  6.239e-01   2.109 0.034925 *
factor(key)2:valence             -5.154e-01  5.654e-01  -0.912 0.361951
factor(key)3:valence             -1.226e+00  7.894e-01  -1.553 0.120409
factor(key)4:valence             -5.308e-01  6.265e-01  -0.847 0.396818
factor(key)5:valence             -7.381e-02  5.980e-01  -0.123 0.901768
factor(key)6:valence              7.618e-03  7.239e-01   0.011 0.991603
factor(key)7:valence              2.848e-01  5.499e-01   0.518 0.604483
factor(key)8:valence             -8.212e-01  6.630e-01  -1.239 0.215506
factor(key)9:valence              1.669e-01  5.813e-01   0.287 0.773976
factor(key)10:valence            -6.984e-01  6.424e-01  -1.087 0.276998
factor(key)11:valence             1.695e+00  6.707e-01   2.528 0.011475 *
speechiness:year                 -2.170e-01  1.024e-02 -21.191  < 2e-16 ***
tempo:year                       -1.810e-04  4.362e-05  -4.149 3.34e-05 ***
tempo:liveness                   -1.272e-02  5.095e-03  -2.497 0.012523 *
tempo:valence                     1.690e-02  4.477e-03   3.775 0.000160 ***
year:liveness                    -3.122e-02  8.275e-03  -3.773 0.000161 ***
year:factor(mode)1               -1.648e-02  2.855e-03  -5.773 7.81e-09 ***
liveness:valence                  3.683e+00  6.466e-01   5.696 1.23e-08 ***
liveness:factor(mode)1            6.628e-01  3.354e-01   1.976 0.048120 *
valence:factor(mode)1            -7.544e-01  2.481e-01  -3.041 0.002360 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 10.61 on 170503 degrees of freedom
Multiple R-squared:  0.7641,    Adjusted R-squared:  0.7639
F-statistic:  3707 on 149 and 170503 DF,  p-value: < 2.2e-16
```

We removed the interaction terms danceability*speechiness, key*mode, speechiness*mode and tempo*mode from our model. Our final interaction model has an RMSE of 10.61, and an R^2_{adj} of 0.7639. An improvement on our first order model by 0.22 and 0.0103, respectively.

Higher order terms

In order to assess the significance of higher order terms, we added single higher order terms and performed t-tests. Terms that have a p-value below 0.05 will reject the null hypothesis stating that they are significant and should be kept in our model. For example, the figure below shows the model with the first higher order term: acousticness^2.

```
Call:
lm(formula = popularity ~ acousticness + I(acousticness^2) +
    danceability + energy + factor(explicit) + instrumentalness +
    factor(key) + speechiness + tempo + year + liveness + valence +
    factor(mode) + acousticness * danceability + acousticness *
    energy + acousticness * factor(explicit) + acousticness *
    factor(key) + acousticness * tempo + acousticness * year +
    acousticness * liveness + acousticness * valence + danceability *
    factor(explicit) + danceability * instrumentalness + danceability *
    factor(key) + danceability * tempo + danceability * year +
    danceability * valence + energy * factor(explicit) + energy *
    instrumentalness + energy * factor(key) + energy * speechiness +
    energy * tempo + energy * liveness + energy * valence + energy *
    factor(mode) + factor(explicit) * instrumentalness + factor(explicit) *
    factor(key) + factor(explicit) * speechiness + factor(explicit) *
    tempo + factor(explicit) * year + factor(explicit) * valence +
    factor(explicit) * factor(mode) + instrumentalness * factor(key) +
    instrumentalness * tempo + instrumentalness * year + instrumentalness *
    liveness + instrumentalness * valence + factor(key) * speechiness +
    factor(key) * year + factor(key) * valence + speechiness *
    year + tempo * year + tempo * liveness + tempo * valence +
    year * liveness + year * factor(mode) + liveness * valence +
    liveness * factor(mode) + valence * factor(mode), data = spotify_data)

Residuals:
    Min      1Q  Median      3Q     Max
-65.764  -7.009  -1.344   5.563  67.647

Coefficients:
                                Estimate Std. Error t value Pr(>|t|)
(Intercept)                   -1.435e+03  1.841e+01 -77.938  < 2e-16 ***
acousticness                  -2.264e+02  9.025e+00 -25.090  < 2e-16 ***
I(acousticness^2)             -1.115e+01  4.618e-01 -24.141  < 2e-16 ***
danceability                   3.941e+02  1.687e+01  23.368  < 2e-16 ***
```

After evaluating the significance of each higher order term one-by-one, we arrived at the following model.

```r
spotify_higherorder =  lm(popularity ~ acousticness + I(acousticness^2)+ I(acousticness^3)+ I(acousticness^4)
+ danceability + energy + I(energy^2)+ I(energy^3) + factor(explicit) + instrumentalness+
I(instrumentalness^2)+ I(instrumentalness^3)+ I(instrumentalness^4) + factor(key) + speechiness+
I(speechiness^2)+ tempo + I(tempo^2)+ I(tempo^3)+ I(tempo^4)+ I(tempo^5) + year+I(year^2)+I(year^3) + liveness
+ valence  + factor(mode) + acousticness*danceability + acousticness*energy + acousticness*factor(explicit) +
acousticness*factor(key) + acousticness*tempo + acousticness*year + acousticness*liveness  +
acousticness*valence + danceability*factor(explicit) + danceability*instrumentalness +
danceability*factor(key) + danceability*tempo + danceability*year  + danceability*valence +
energy*factor(explicit) + energy*instrumentalness + energy*factor(key) + energy*speechiness + energy*tempo +
energy*liveness  + energy*valence + energy*factor(mode) + factor(explicit)*instrumentalness +
factor(explicit)*factor(key) + factor(explicit)*speechiness + factor(explicit)*tempo + factor(explicit)*year +
factor(explicit)*valence  + factor(explicit)*factor(mode) + instrumentalness*factor(key) +
instrumentalness*tempo + instrumentalness*year + instrumentalness*liveness + instrumentalness*valence +
factor(key)*speechiness + factor(key)*year + factor(key)*valence + speechiness*year + tempo*year +
tempo*liveness + tempo*valence + year*liveness + year*factor(mode) + liveness*valence + liveness*factor(mode)
+ valence*factor(mode), data = spotify_data)
summary(spotify_higherorder)
```

```
Call:
lm(formula = popularity ~ acousticness + I(acousticness^2) +
    I(acousticness^3) + I(acousticness^4) + danceability + energy +
    I(energy^2) + I(energy^3) + factor(explicit) + instrumentalness +
    I(instrumentalness^2) + I(instrumentalness^3) + I(instrumentalness^4) +
    factor(key) + speechiness + I(speechiness^2) + tempo + I(tempo^2) +
    I(tempo^3) + I(tempo^4) + I(tempo^5) + year + I(year^2) +
    I(year^3) + liveness + valence + factor(mode) + acousticness *
    danceability + acousticness * energy + acousticness * factor(explicit) +
    acousticness * factor(key) + acousticness * tempo + acousticness *
    year + acousticness * liveness + acousticness * valence +
    danceability * factor(explicit) + danceability * instrumentalness +
    danceability * factor(key) + danceability * tempo + danceability *
    year + danceability * valence + energy * factor(explicit) +
    energy * instrumentalness + energy * factor(key) + energy *
    speechiness + energy * tempo + energy * liveness + energy *
    valence + energy * factor(mode) + factor(explicit) * instrumentalness +
    factor(explicit) * factor(key) + factor(explicit) * speechiness +
    factor(explicit) * tempo + factor(explicit) * year + factor(explicit) *
    valence + factor(explicit) * factor(mode) + instrumentalness *
    factor(key) + instrumentalness * tempo + instrumentalness *
    year + instrumentalness * liveness + instrumentalness * valence +
    factor(key) * speechiness + factor(key) * year + factor(key) *
    valence + speechiness * year + tempo * year + tempo * liveness +
    tempo * valence + year * liveness + year * factor(mode) +
    liveness * valence + liveness * factor(mode) + valence *
    factor(mode), data = spotify_data)

Residuals:
    Min      1Q  Median      3Q     Max
-64.917  -6.737  -1.344   5.137  67.789

Coefficients:
                                Estimate Std. Error t value Pr(>|t|)
(Intercept)                    5.383e+05  1.336e+04  40.283  < 2e-16 ***
acousticness                  -2.619e+02  1.026e+01 -25.517  < 2e-16 ***
I(acousticness^2)             -7.427e+01  6.398e+00 -11.609  < 2e-16 ***
I(acousticness^3)              1.188e+02  9.939e+00  11.956  < 2e-16 ***
I(acousticness^4)             -6.693e+01  4.974e+00 -13.454  < 2e-16 ***
danceability                   2.930e+02  1.698e+01  17.253  < 2e-16 ***
energy                         2.323e+00  1.582e+00   1.469 0.141959
I(energy^2)                    1.225e+01  2.694e+00   4.547 5.46e-06 ***
I(energy^3)                   -1.269e+01  1.685e+00  -7.533 5.00e-14 ***
factor(explicit)1             -4.731e+02  1.576e+01 -30.019  < 2e-16 ***
instrumentalness               8.341e+01  8.780e+00   9.500  < 2e-16 ***
I(instrumentalness^2)          6.584e+01  9.342e+00   7.048 1.83e-12 ***
I(instrumentalness^3)         -1.021e+02  1.600e+01  -6.379 1.79e-10 ***
I(instrumentalness^4)          5.156e+01  8.564e+00   6.020 1.74e-09 ***
```

factor(key)1	-5.441e+01	1.277e+01	-4.260	2.04e-05	***
factor(key)2	-6.640e+00	1.167e+01	-0.569	0.569280	
factor(key)3	3.038e+01	1.531e+01	1.985	0.047178	*
factor(key)4	2.138e+01	1.309e+01	1.634	0.102253	
factor(key)5	-3.303e+01	1.193e+01	-2.768	0.005634	**
factor(key)6	-2.424e+01	1.481e+01	-1.636	0.101775	
factor(key)7	4.721e+00	1.125e+01	0.420	0.674799	
factor(key)8	-4.743e+01	1.362e+01	-3.481	0.000499	***
factor(key)9	1.350e+01	1.204e+01	1.122	0.262070	
factor(key)10	-1.571e+01	1.316e+01	-1.194	0.232413	
factor(key)11	6.703e+00	1.421e+01	0.472	0.637229	
speechiness	2.007e+02	2.315e+01	8.670	< 2e-16	***
I(speechiness^2)	7.017e+00	9.425e-01	7.444	9.78e-14	***
tempo	3.431e-01	1.034e-01	3.318	0.000907	***
I(tempo^2)	3.225e-03	1.204e-03	2.678	0.007415	**
I(tempo^3)	-3.991e-05	1.272e-05	-3.138	0.001701	**
I(tempo^4)	2.109e-07	6.078e-08	3.469	0.000522	***
I(tempo^5)	-3.947e-10	1.077e-10	-3.666	0.000247	***
year	-8.182e+02	2.032e+01	-40.263	< 2e-16	***
I(year^2)	4.142e-01	1.030e-02	40.209	< 2e-16	***
I(year^3)	-6.982e-05	1.740e-06	-40.119	< 2e-16	***
liveness	5.283e+01	1.649e+01	3.205	0.001353	**
valence	1.010e+00	9.798e-01	1.031	0.302715	
factor(mode)1	3.999e+01	5.571e+00	7.179	7.05e-13	***
acousticness:danceability	7.113e+00	6.702e-01	10.613	< 2e-16	***
acousticness:energy	-7.020e+00	7.505e-01	-9.354	< 2e-16	***
acousticness:factor(explicit)1	-1.599e+00	4.736e-01	-3.377	0.000733	***
acousticness:factor(key)1	-6.611e-01	5.356e-01	-1.234	0.217132	
acousticness:factor(key)2	-7.865e-01	4.739e-01	-1.660	0.096979	.
acousticness:factor(key)3	-1.328e+00	7.027e-01	-1.889	0.058835	.
acousticness:factor(key)4	-1.063e+00	5.323e-01	-1.997	0.045880	*
acousticness:factor(key)5	-8.178e-01	4.989e-01	-1.639	0.101171	
acousticness:factor(key)6	-1.995e+00	6.058e-01	-3.292	0.000994	***
acousticness:factor(key)7	-1.226e+00	4.609e-01	-2.661	0.007798	**
acousticness:factor(key)8	3.025e-01	5.846e-01	0.517	0.604834	
acousticness:factor(key)9	-1.064e+00	4.806e-01	-2.214	0.026815	*
acousticness:factor(key)10	-1.059e+00	5.570e-01	-1.902	0.057188	.
acousticness:factor(key)11	-1.675e+00	5.526e-01	-3.032	0.002432	**
acousticness:tempo	1.119e-02	3.860e-03	2.899	0.003744	**
acousticness:year	1.422e-01	5.070e-03	28.056	< 2e-16	***
acousticness:liveness	-1.293e+00	6.488e-01	-1.992	0.046332	*
acousticness:valence	-6.063e+00	5.275e-01	-11.495	< 2e-16	***
danceability:factor(explicit)1	5.058e+00	7.580e-01	6.673	2.51e-11	***
danceability:instrumentalness	-1.050e+01	6.332e-01	-16.578	< 2e-16	***
danceability:factor(key)1	-2.486e+00	8.848e-01	-2.809	0.004966	**
danceability:factor(key)2	-7.733e-02	8.132e-01	-0.095	0.924242	
danceability:factor(key)3	8.395e-01	1.140e+00	0.736	0.461510	
danceability:factor(key)4	7.354e-01	9.053e-01	0.812	0.416651	
danceability:factor(key)5	3.015e-01	8.576e-01	0.352	0.725200	
danceability:factor(key)6	-4.100e-01	1.030e+00	-0.398	0.690616	
danceability:factor(key)7	-1.620e+00	7.930e-01	-2.042	0.041140	*
danceability:factor(key)8	7.218e-01	9.707e-01	0.744	0.457128	
danceability:factor(key)9	-8.568e-01	8.426e-01	-1.017	0.309241	
danceability:factor(key)10	3.060e-01	9.297e-01	0.329	0.742012	
danceability:factor(key)11	-2.725e+00	9.607e-01	-2.837	0.004561	**
danceability:tempo	2.745e-03	7.220e-03	0.380	0.703786	
danceability:year	-1.467e-01	8.464e-03	-17.329	< 2e-16	***
danceability:valence	-2.216e+00	6.545e-01	-3.387	0.000708	***
energy:factor(explicit)1	-7.238e+00	6.709e-01	-10.788	< 2e-16	***
energy:instrumentalness	-6.406e+00	4.477e-01	-14.308	< 2e-16	***
energy:factor(key)1	-1.796e+00	7.340e-01	-2.447	0.014422	*
energy:factor(key)2	-4.749e-01	6.595e-01	-0.720	0.471460	
energy:factor(key)3	7.337e-01	9.397e-01	0.781	0.434946	
energy:factor(key)4	-3.879e-01	7.441e-01	-0.521	0.602162	
energy:factor(key)5	-1.128e+00	6.993e-01	-1.614	0.106612	
energy:factor(key)6	-1.955e+00	8.434e-01	-2.318	0.020448	*
energy:factor(key)7	-2.073e+00	6.487e-01	-3.196	0.001392	**
energy:factor(key)8	-2.192e-01	7.974e-01	-0.275	0.783436	
energy:factor(key)9	-9.619e-01	6.793e-01	-1.416	0.156726	

```
factor(explicit)1:factor(key)10    -1.251e+00  5.468e-01  -2.289 0.022095 *
factor(explicit)1:factor(key)11    -4.008e-01  5.038e-01  -0.795 0.426334
factor(explicit)1:speechiness       7.739e+00  7.239e-01  10.691  < 2e-16 ***
factor(explicit)1:tempo             1.876e-02  3.332e-03   5.628 1.82e-08 ***
factor(explicit)1:year              2.372e-01  7.897e-03  30.029  < 2e-16 ***
factor(explicit)1:valence          -1.487e+00  5.109e-01  -2.911 0.003601 **
factor(explicit)1:factor(mode)1     7.151e-01  2.175e-01   3.287 0.001013 **
instrumentalness:factor(key)1      -1.979e-01  4.176e-01  -0.474 0.635649
instrumentalness:factor(key)2       5.313e-01  3.673e-01   1.446 0.148070
instrumentalness:factor(key)3      -6.845e-01  4.596e-01  -1.489 0.136388
instrumentalness:factor(key)4      -2.606e-01  4.148e-01  -0.628 0.529920
instrumentalness:factor(key)5       4.188e-01  3.688e-01   1.135 0.256179
instrumentalness:factor(key)6      -2.340e-01  4.929e-01  -0.475 0.635046
instrumentalness:factor(key)7      -3.818e-02  3.566e-01  -0.107 0.914732
instrumentalness:factor(key)8       1.914e-01  4.250e-01   0.450 0.652546
instrumentalness:factor(key)9       3.134e-01  3.852e-01   0.814 0.415901
instrumentalness:factor(key)10     -6.980e-01  4.073e-01  -1.714 0.086620 .
instrumentalness:factor(key)11     -7.671e-01  4.663e-01  -1.645 0.099940 .
instrumentalness:tempo             -1.514e-02  2.997e-03  -5.051 4.40e-07 ***
instrumentalness:year              -4.849e-02  4.352e-03 -11.143  < 2e-16 ***
instrumentalness:liveness           2.165e+00  5.636e-01   3.841 0.000123 ***
instrumentalness:valence            6.969e+00  4.417e-01  15.778  < 2e-16 ***
factor(key)1:speechiness            1.902e+00  8.066e-01   2.359 0.018346 *
factor(key)2:speechiness            1.294e+00  8.868e-01   1.459 0.144478
factor(key)3:speechiness            4.868e+00  1.133e+00   4.297 1.73e-05 ***
factor(key)4:speechiness            3.521e-01  9.955e-01   0.354 0.723607
factor(key)5:speechiness            1.160e+00  8.932e-01   1.299 0.194086
factor(key)6:speechiness            1.759e+00  8.949e-01   1.966 0.049333 *
factor(key)7:speechiness            1.355e+00  8.247e-01   1.643 0.100370
factor(key)8:speechiness            1.514e+00  9.671e-01   1.565 0.117578
factor(key)9:speechiness            9.503e-01  8.323e-01   1.142 0.253538
factor(key)10:speechiness           3.825e+00  8.682e-01   4.406 1.05e-05 ***
factor(key)11:speechiness           1.500e+00  8.473e-01   1.771 0.076628 .
factor(key)1:year                   2.851e-02  6.431e-03   4.434 9.28e-06 ***
factor(key)2:year                   3.745e-03  5.876e-03   0.637 0.523904
factor(key)3:year                  -1.514e-02  7.668e-03  -1.974 0.048375 *
factor(key)4:year                  -1.030e-02  6.590e-03  -1.563 0.117940
factor(key)5:year                   1.728e-02  6.005e-03   2.878 0.004002 **
factor(key)6:year                   1.347e-02  7.466e-03   1.804 0.071170 .
factor(key)7:year                  -1.307e-03  5.670e-03  -0.231 0.817625
factor(key)8:year                   2.415e-02  6.858e-03   3.522 0.000428 ***
factor(key)9:year                  -6.092e-03  6.070e-03  -1.004 0.315499
factor(key)10:year                  8.852e-03  6.618e-03   1.338 0.181036
factor(key)11:year                 -2.302e-03  7.176e-03  -0.321 0.748347
factor(key)1:valence                1.424e+00  6.198e-01   2.298 0.021580 *
factor(key)2:valence               -5.676e-01  5.616e-01  -1.011 0.312187
factor(key)3:valence               -1.230e+00  7.841e-01  -1.568 0.116875
factor(key)4:valence               -5.449e-01  6.223e-01  -0.876 0.381214
factor(key)5:valence               -3.085e-01  5.940e-01  -0.519 0.603482
factor(key)6:valence               -1.079e-02  7.192e-01  -0.015 0.988030
factor(key)7:valence                2.180e-01  5.462e-01   0.399 0.689835
factor(key)8:valence               -9.915e-01  6.587e-01  -1.505 0.132284
factor(key)9:valence                7.510e-02  5.775e-01   0.130 0.896523
factor(key)10:valence              -1.168e+00  6.383e-01  -1.830 0.067316 .
factor(key)11:valence               1.782e+00  6.661e-01   2.675 0.007473 **
speechiness:year                   -1.140e-01  1.175e-02  -9.696  < 2e-16 ***
tempo:year                         -2.277e-04  4.357e-05  -5.225 1.74e-07 ***
tempo:liveness                     -1.028e-02  5.081e-03  -2.024 0.042978 *
tempo:valence                       8.579e-04  4.681e-03   0.183 0.854590
year:liveness                      -2.851e-02  8.261e-03  -3.451 0.000558 ***
year:factor(mode)1                 -2.062e-02  2.838e-03  -7.265 3.75e-13 ***
liveness:valence                    3.547e+00  6.456e-01   5.495 3.92e-08 ***
liveness:factor(mode)1              5.549e-01  3.331e-01   1.666 0.095727 .
valence:factor(mode)1              -8.344e-01  2.464e-01  -3.386 0.000710 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 10.53 on 170489 degrees of freedom
Multiple R-squared:  0.7674,    Adjusted R-squared:  0.7672
F-statistic:  3451 on 163 and 170489 DF,  p-value: < 2.2e-16
```

This higher order model contains a total of 15 significant higher order terms. As we expected, this model makes more improvements on the interaction model and has an RMSE of 10.53 and an R^2_{adj} of 0.7672. These values represent improvements on the interaction model by 0.08 and 0.0033, respectively.

Log Model

```{r}
bestfit = lm(popularity ~ acousticness + I(acousticness^2)+ I(acousticness^3)+ I(acousticness^4) +
danceability + energy + I(energy^2)+ I(energy^3) + factor(explicit) + instrumentalness+ I(instrumentalness^2)+
I(instrumentalness^3)+ I(instrumentalness^4) + factor(key) + speechiness+ I(speechiness^2)+ tempo +
I(tempo^2)+ I(tempo^3)+ I(tempo^4)+ I(tempo^5) + log(year)+I(year^2)+I(year^3) + liveness + valence  +
factor(mode)
                      + acousticness*danceability + acousticness*energy + acousticness*factor(explicit) +
acousticness*factor(key) + acousticness*tempo + acousticness*year + acousticness*liveness
                      + acousticness*valence + danceability*factor(explicit) + danceability*instrumentalness
+ danceability*factor(key) + danceability*tempo + danceability*year
                      + danceability*valence + energy*factor(explicit) + energy*instrumentalness +
energy*factor(key) + energy*speechiness + energy*tempo + energy*liveness
                      + energy*valence + energy*factor(mode) + factor(explicit)*instrumentalness +
factor(explicit)*factor(key) + factor(explicit)*speechiness + factor(explicit)*tempo
                      + factor(explicit)*year + factor(explicit)*valence
                      + factor(explicit)*factor(mode) + instrumentalness*factor(key) + instrumentalness*tempo
+ instrumentalness*year + instrumentalness*liveness + instrumentalness*valence
                      + factor(key)*speechiness + factor(key)*year + factor(key)*valence + speechiness*year +
tempo*year + tempo*liveness
                      + tempo*valence
                      + year*liveness + year*factor(mode) + liveness*valence + liveness*factor(mode) +
valence*factor(mode), data = spotify_data)

summary(bestfit)
```

```
Call:
lm(formula = popularity ~ acousticness + I(acousticness^2) +
    I(acousticness^3) + I(acousticness^4) + danceability + energy +
    I(energy^2) + I(energy^3) + factor(explicit) + instrumentalness +
    I(instrumentalness^2) + I(instrumentalness^3) + I(instrumentalness^4) +
    factor(key) + speechiness + I(speechiness^2) + tempo + I(tempo^2) +
    I(tempo^3) + I(tempo^4) + I(tempo^5) + log(year) + I(year^2) +
    I(year^3) + liveness + valence + factor(mode) + acousticness *
    danceability + acousticness * energy + acousticness * factor(explicit) +
    acousticness * factor(key) + acousticness * tempo + acousticness *
    year + acousticness * liveness + acousticness * valence +
    danceability * factor(explicit) + danceability * instrumentalness +
    danceability * factor(key) + danceability * tempo + danceability *
    year + danceability * valence + energy * factor(explicit) +
    energy * instrumentalness + energy * factor(key) + energy *
    speechiness + energy * tempo + energy * liveness + energy *
    valence + energy * factor(mode) + factor(explicit) * instrumentalness +
    factor(explicit) * factor(key) + factor(explicit) * speechiness +
    factor(explicit) * tempo + factor(explicit) * year + factor(explicit) *
    valence + factor(explicit) * factor(mode) + instrumentalness *
    factor(key) + instrumentalness * tempo + instrumentalness *
    year + instrumentalness * liveness + instrumentalness * valence +
    factor(key) * speechiness + factor(key) * year + factor(key) *
    valence + speechiness * year + tempo * year + tempo * liveness +
    tempo * valence + year * liveness + year * factor(mode) +
    liveness * valence + liveness * factor(mode) + valence *
    factor(mode), data = spotify_data)

Residuals:
    Min      1Q  Median      3Q     Max
-68.037  -6.252  -1.149   4.661  73.328
```

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	2.270e+09	2.348e+07	96.644	< 2e-16	***
acousticness	-1.027e+02	1.013e+01	-10.142	< 2e-16	***
I(acousticness^2)	-6.389e+01	6.230e+00	-10.255	< 2e-16	***
I(acousticness^3)	1.031e+02	9.679e+00	10.654	< 2e-16	***
I(acousticness^4)	-5.728e+01	4.845e+00	-11.824	< 2e-16	***
danceability	2.538e+02	1.654e+01	15.341	< 2e-16	***
energy	3.207e+00	1.540e+00	2.082	0.037345	*
I(energy^2)	9.251e+00	2.623e+00	3.527	0.000421	***
I(energy^3)	-9.190e+00	1.641e+00	-5.600	2.15e-08	***
factor(explicit)1	-2.708e+02	1.549e+01	-17.484	< 2e-16	***
instrumentalness	7.646e+01	8.549e+00	8.944	< 2e-16	***
I(instrumentalness^2)	6.834e+01	9.097e+00	7.512	5.84e-14	***
I(instrumentalness^3)	-1.023e+02	1.558e+01	-6.565	5.21e-11	***
I(instrumentalness^4)	4.923e+01	8.339e+00	5.904	3.56e-09	***
factor(key)1	-4.142e+01	1.244e+01	-3.331	0.000867	***
factor(key)2	-9.694e+00	1.136e+01	-0.853	0.393495	
factor(key)3	2.786e+01	1.490e+01	1.870	0.061522	.
factor(key)4	3.304e+00	1.274e+01	0.259	0.795430	
factor(key)5	-2.792e+01	1.162e+01	-2.403	0.016256	*
factor(key)6	-2.365e+01	1.442e+01	-1.640	0.101039	
factor(key)7	7.573e+00	1.096e+01	0.691	0.489410	
factor(key)8	-3.504e+01	1.327e+01	-2.641	0.008263	**
factor(key)9	-3.464e+00	1.172e+01	-0.296	0.767573	
factor(key)10	-1.088e+01	1.281e+01	-0.850	0.395594	
factor(key)11	1.200e+00	1.384e+01	0.087	0.930884	
speechiness	1.231e+02	2.256e+01	5.455	4.91e-08	***
I(speechiness^2)	8.046e+00	9.178e-01	8.767	< 2e-16	***
tempo	2.185e-01	1.007e-01	2.171	0.029959	*
I(tempo^2)	3.437e-03	1.173e-03	2.931	0.003379	**
I(tempo^3)	-4.549e-05	1.238e-05	-3.673	0.000239	***
I(tempo^4)	2.502e-07	5.919e-08	4.227	2.37e-05	***
I(tempo^5)	-4.817e-10	1.048e-10	-4.594	4.34e-06	***
log(year)	-3.944e+08	4.081e+06	-96.622	< 2e-16	***
I(year^2)	-1.516e+02	1.573e+00	-96.356	< 2e-16	***
I(year^3)	1.705e-02	1.772e-04	96.223	< 2e-16	***
liveness	5.122e+01	1.605e+01	3.191	0.001420	**
valence	-6.079e-01	9.542e-01	-0.637	0.524083	
factor(mode)1	2.225e+01	5.427e+00	4.099	4.15e-05	***
year	5.989e+05	6.207e+03	96.489	< 2e-16	***
acousticness:danceability	6.377e+00	6.526e-01	9.772	< 2e-16	***
acousticness:energy	-5.233e+00	7.310e-01	-7.159	8.15e-13	***
acousticness:factor(explicit)1	-1.809e+00	4.611e-01	-3.923	8.75e-05	***
acousticness:factor(key)1	-5.868e-01	5.215e-01	-1.125	0.260500	
acousticness:factor(key)2	-5.353e-01	4.614e-01	-1.160	0.245997	
acousticness:factor(key)3	-1.986e+00	6.842e-01	-2.902	0.003702	**
acousticness:factor(key)4	-4.732e-01	5.183e-01	-0.913	0.361295	
acousticness:factor(key)5	-8.658e-01	4.858e-01	-1.782	0.074691	.
acousticness:factor(key)6	-1.856e+00	5.899e-01	-3.146	0.001656	**
acousticness:factor(key)7	-1.187e+00	4.488e-01	-2.644	0.008183	**
acousticness:factor(key)8	-1.106e-01	5.692e-01	-0.194	0.845884	
acousticness:factor(key)9	-3.402e-01	4.681e-01	-0.727	0.467392	
acousticness:factor(key)10	-1.179e+00	5.424e-01	-2.173	0.029788	*
acousticness:factor(key)11	-1.388e+00	5.380e-01	-2.580	0.009890	**
acousticness:tempo	1.459e-02	3.759e-03	3.882	0.000104	***
acousticness:year	6.012e-02	5.009e-03	12.003	< 2e-16	***
acousticness:liveness	-1.726e+00	6.317e-01	-2.732	0.006287	**
acousticness:valence	-5.293e+00	5.137e-01	-10.305	< 2e-16	***

```
factor(key)1:speechiness           1.440e+00  7.854e-01   1.834 0.066706 .
factor(key)2:speechiness           1.170e+00  8.635e-01   1.355 0.175412
factor(key)3:speechiness           5.642e+00  1.103e+00   5.114 3.16e-07 ***
factor(key)4:speechiness           4.058e-01  9.693e-01   0.419 0.675452
factor(key)5:speechiness           5.145e-02  8.697e-01   0.059 0.952825
factor(key)6:speechiness           1.050e+00  8.714e-01   1.205 0.228304
factor(key)7:speechiness           1.184e+00  8.030e-01   1.474 0.140446
factor(key)8:speechiness           2.229e-01  9.418e-01   0.237 0.812896
factor(key)9:speechiness          -1.249e-01  8.105e-01  -0.154 0.877512
factor(key)10:speechiness          3.853e+00  8.454e-01   4.558 5.17e-06 ***
factor(key)11:speechiness          8.645e-01  8.250e-01   1.048 0.294716
factor(key)1:year                  2.179e-02  6.262e-03   3.479 0.000503 ***
factor(key)2:year                  5.035e-03  5.721e-03   0.880 0.378844
factor(key)3:year                 -1.360e-02  7.467e-03  -1.822 0.068489 .
factor(key)4:year                 -1.546e-03  6.418e-03  -0.241 0.809637
factor(key)5:year                  1.461e-02  5.847e-03   2.499 0.012461 *
factor(key)6:year                  1.305e-02  7.269e-03   1.795 0.072710 .
factor(key)7:year                 -2.896e-03  5.521e-03  -0.525 0.599852
factor(key)8:year                  1.805e-02  6.678e-03   2.703 0.006870 **
factor(key)9:year                  2.072e-03  5.911e-03   0.350 0.725968
factor(key)10:year                 6.418e-03  6.444e-03   0.996 0.319247
factor(key)11:year                 3.971e-04  6.987e-03   0.057 0.954678
factor(key)1:valence               1.360e+00  6.035e-01   2.254 0.024212 *
factor(key)2:valence              -9.666e-01  5.468e-01  -1.768 0.077129 .
factor(key)3:valence              -9.109e-01  7.635e-01  -1.193 0.232846
factor(key)4:valence              -9.232e-01  6.060e-01  -1.523 0.127646
factor(key)5:valence              -3.979e-01  5.784e-01  -0.688 0.491490
factor(key)6:valence               1.866e-01  7.003e-01   0.266 0.789928
factor(key)7:valence               5.241e-02  5.318e-01   0.099 0.921491
factor(key)8:valence              -6.014e-01  6.414e-01  -0.938 0.348488
factor(key)9:valence               1.264e-03  5.623e-01   0.002 0.998207
factor(key)10:valence             -1.022e+00  6.215e-01  -1.645 0.100019
factor(key)11:valence              2.163e+00  6.486e-01   3.334 0.000855 ***
speechiness:year                  -7.271e-02  1.145e-02  -6.349 2.17e-10 ***
tempo:year                        -1.604e-04  4.243e-05  -3.781 0.000156 ***
tempo:liveness                    -9.812e-03  4.948e-03  -1.983 0.047362 *
tempo:valence                      5.310e-03  4.558e-03   1.165 0.244022
liveness:year                     -2.711e-02  8.044e-03  -3.370 0.000752 ***
factor(mode)1:year                -1.152e-02  2.765e-03  -4.166 3.11e-05 ***
liveness:valence                   4.602e+00  6.287e-01   7.320 2.49e-13 ***
liveness:factor(mode)1             2.252e-01  3.244e-01   0.694 0.487404
valence:factor(mode)1             -8.877e-01  2.400e-01  -3.699 0.000216 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 10.26 on 170488 degrees of freedom
Multiple R-squared:  0.7795,    Adjusted R-squared:  0.7792
F-statistic:  3674 on 164 and 170488 DF,  p-value: < 2.2e-16
```

We chose to log the year variable as it had the largest absolute t-value in our stepwise regression and was determined as the best one-variable predictor of Y. After including the transformation, it did seem to help, however not enough to make our residuals normal, or make our variance in our error term equal. We arrive at our final model with an RMSE of 10.26, and an R^2_{adj} of 0.7792. These values represent an improvement on the higher order model of 0.27 and 0.012, respectively.

Our Best Fit Model

Our Best Fit Model:

$$
\begin{aligned}
\widehat{popularity} = {} & \beta_0 + \beta_1 acousticness + \beta_2 acousticness^2 + \beta_3 acousticness^3 + \beta_4 acousticness^4 \\
& + \beta_5 dancability + \beta_6 enegry + \beta_7 enegry^2 + \beta_8 enegry^3 + \beta_9 instrumentals \\
& + \beta_{10} instrumentals^2 + \beta_{11} instrumentals^3 + \beta_{12} instrumentals^4 + \beta_{13} speachness \\
& + \beta_{14} speachness^2 + \beta_{15} tempo + \beta_{16} tempo^2 + \beta_{17} tempo^3 + \beta_{18} tempo^4 + \beta_{19} tempo^5 \\
& + \beta_{20} year + \beta_{21} log(year) + \beta_{22} year^2 + \beta_{23} year^3 + \beta_{24} liveness + \beta_{25} valence \\
& + \beta_{26:27} explicit + \beta_{28:39} key + \beta_{40:41} mode \\
& + \beta_{42:60} acousticness(dancibility + energy + explicit + key + tempo + year + liveness + valence) \\
& + \beta_{61:77} dancability(expicit + instrumentals + key + tempo + year + valence) \\
& + \beta_{78:95} energy(instrumentalness + key + speachiness + tempo + liveness + valence + mode) \\
& + \beta_{96:113} explict(instrumentalness + key + speachiness + tempo + year + valence + mode) \\
& + \beta_{114:128} instrumentals(key + tempo + year + liveness + valence) \\
& + \beta_{129:162} key(speachness + year + valence) \\
& + \beta_{163} speachness * year \\
& + \beta_{164:167} tempo(year + liveness + valence) \\
& + \beta_{168:169} year(liveness + mode) \\
& + \beta_{170:171} liveness(valence + mode) \\
& + \beta_{172} valence * mode
\end{aligned}
$$

Why is this our Best Fit Model?

After running various tests and applying higher order terms, and interaction terms, we have decided that the model above was indeed the best model to determine the popularity of a song. These terms were all significant in our model, and also provided the lowest RMSE, along with the highest R^2_{adj} when compared to other models.

From our best fit model, we observe an R^2_{adj} value of 0.7792, and a residual standard error of 10.26. This means that our model accounts for 77.92% of the variance of the popularity of a song and on average, our model would be off by about 10.26 in popularity value.

Assumption Checks

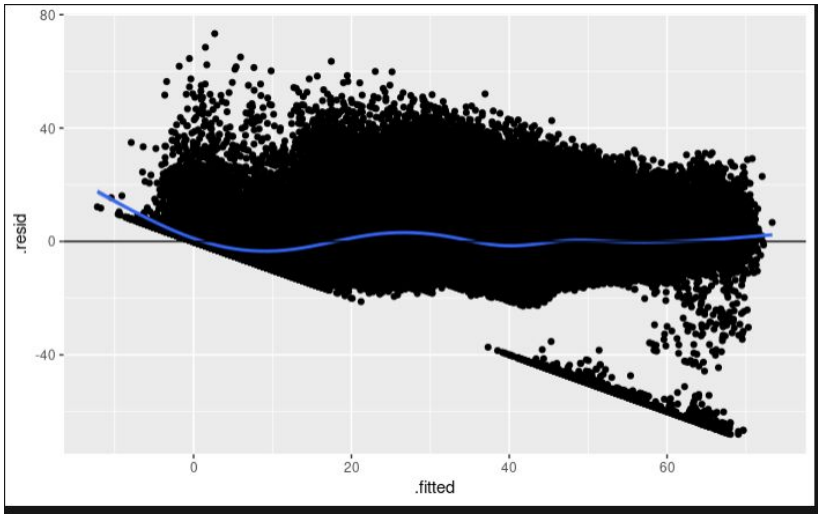
Using our final model, we will be checking the following assumptions:

- Linearity Assumption
- Independence Assumption
- Equal Variance Assumption

- Normality Assumption
- Influential Points and Outliers

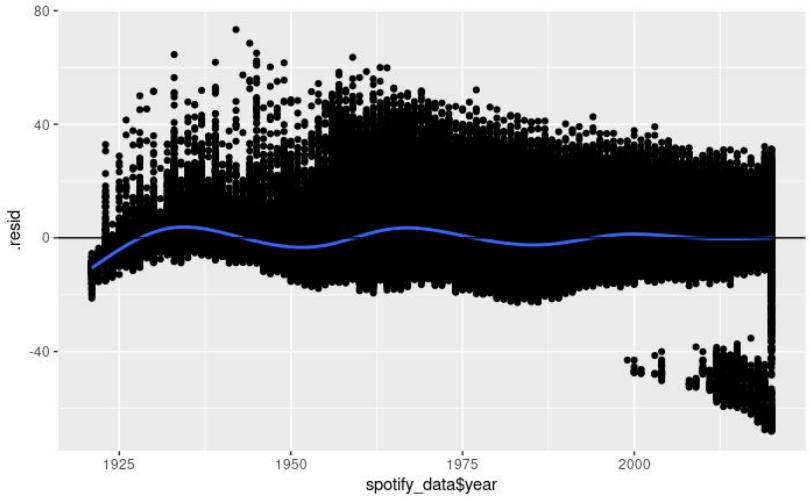
Linearity Assumption

We will check the assumption that there is a straight-line relationship between the predictions and the response variable using the fitted vs residuals plot. From our residual plot below, it does not look like there is a pattern in our data, and our transformations do fit the pattern, which is most prevalent at the left of the chart.



Independence Assumption

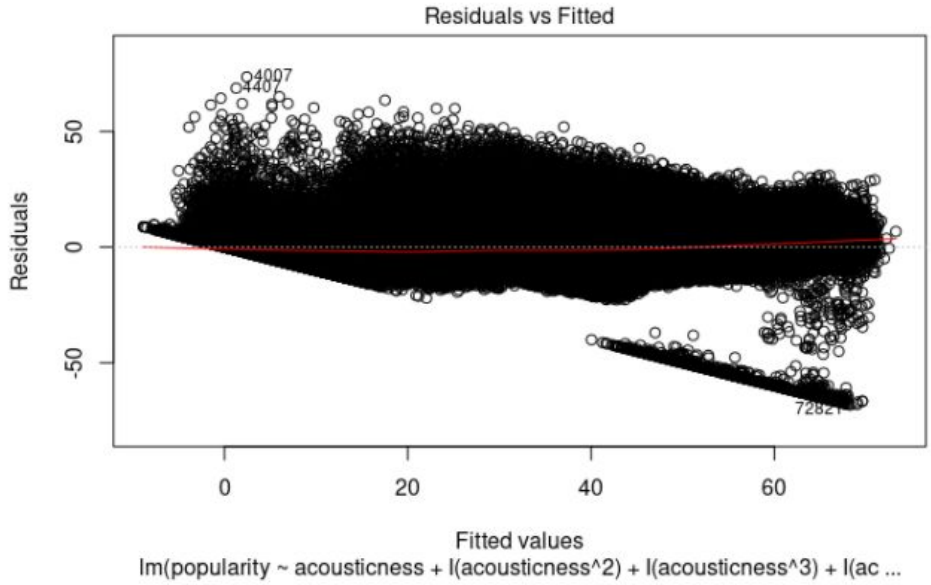
As our data did include a time variable of year, we will check to see if the error terms are uncorrelated. From our graph plotting our residuals over the years, it is hard to see if there are any patterns due to the amount of data we have. There is a clump of data points after the year 2000 that could be worrying.

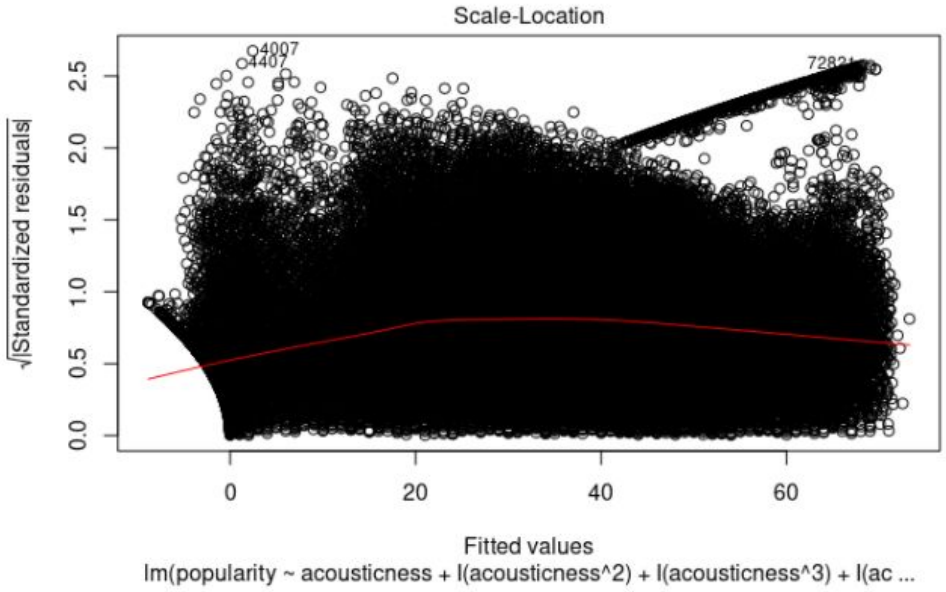


Equal Variance Assumptions

We have checked to see if our models' error terms have a constant variance, so we are not having issues with our response varaible. using a residual plot, a scale location plot, along with the Breusch-Pagan test. With this test, we will be using the following hypothesis:

$$H_0 : \text{heteroscedasticity is not present (homoscedasticity)}$$
$$H_a : \text{heteroscedasticity is present}$$





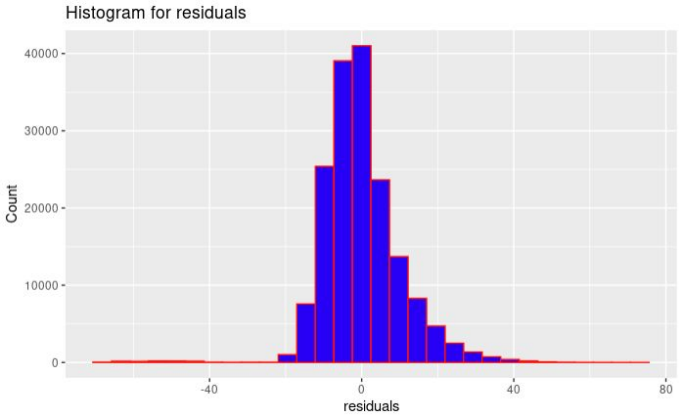
```
        studentized Breusch-Pagan test

data:  bestfit
BP = 9952, df = 163, p-value < 0.00000000000000022
```

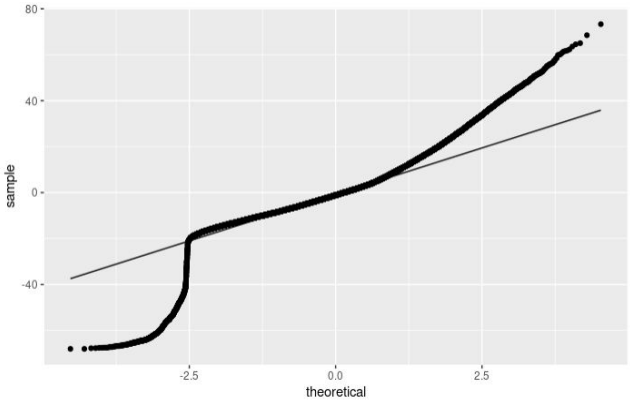
The Residuals vs Fitted plot seems relatively flat, however we see a pattern in the residuals for our Scale-Location plot. This indicates heteroscedasticity and is confirmed by our Breusch-Pagan test. With a p-value of 2.2e-16, which is less than 0.05, we reject the null hypothesis, and evidence suggests that heteroscedasticity does exist in our model.

Normality Assumptions

Our model will also require that the residuals be normally distributed. We will check this assumption using a histogram, a Q-Q plot, and also the Anderson-Darling normality test. We are unable to use the Shapiro-Wilk test as our sample size is over 5000. With this test we will be using the following hypothesis:

$$H_0 : \text{the sample data are significantly normally distributed}$$
$$H_a : \text{the sample data are not significantly normally distributed}$$






```{r}
library(nortest)
ad.test(residuals(bestfit))
```

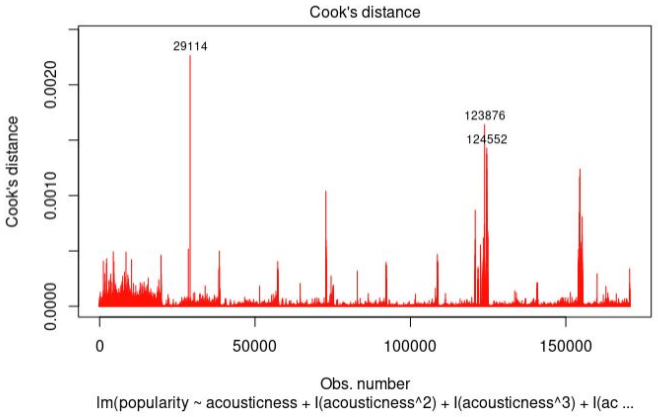
```
        Anderson-Darling normality test

data:  residuals(bestfit)
A = 2422.5, p-value < 0.00000000000000022
```

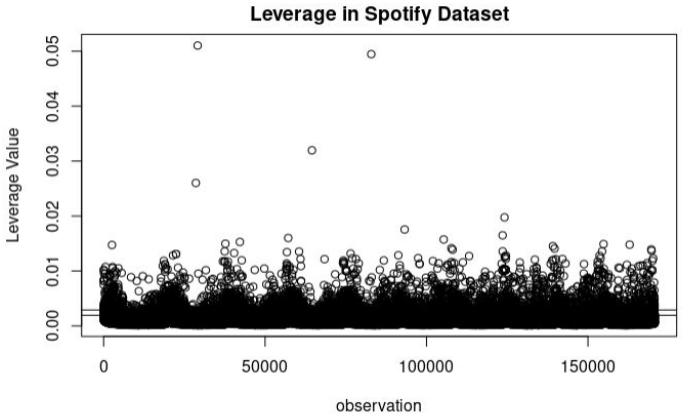
From our graphs, we are showing a pretty good looking histogram of the residuals,however, our Q-Q plot does have an interesting curve at the start that would suggest our residuals are not distributed normally. Using the Anderson-Darling normality test, we observe a p-value of 2.2e-16, which is less than 0.05, we reject the null hypothesis and can suggest that our residuals are not normally distributed.

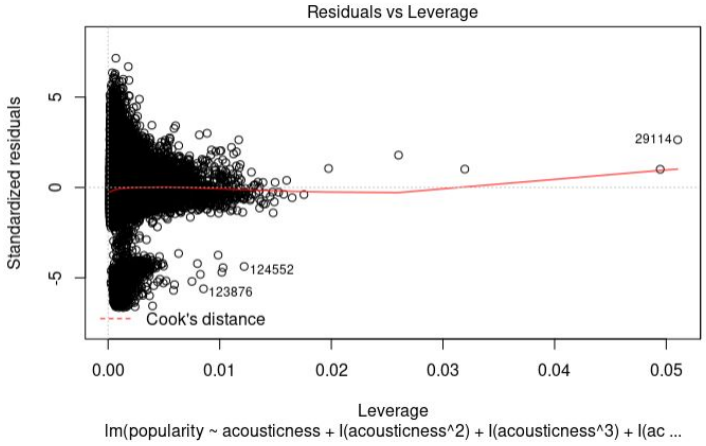
Due to our response variables having non-positive (zeros) entries, we are unable to use the Box-Cox transform to help make our residuals act more normally. Also, due to the large amount of data we have, there seems to be some entries that may be extreme, and cause the strange curves in our Q-Q plot. We had explored ways in which we would be able to transform our variables to provide more normally distributed residuals, however the transformations we found seem to only assist models working with gravity or zero sum trading with foreign countries.

Influential Points and Outliers

Looking to see if there are any influential outliers in our model, we plot our values against Cook's distance to see if there are any influential points. From our graph, it looks like our data is very consistent, as we are unable to see any of the Cook's distance lines in our graph, and there are no influential points that have a disproportionate effect. With having such a high number of data points, it seems that our 2* p/n is a bit skewed, and it may have been higher. We had tried to remove any outliers that were about this 2*p/n threshold, however, our system limited to a maximum of 14k out of the 170k entries. We then decided to use 3*p/n, and saw a drop of 5353 data points from our model.







Interpreting the Effects of Each Independent Variable

After choosing our best fit model, it is interesting to see how the independent variables in our best fit model affect the popularity of a song. We did expect to see many interaction terms as this is the nature of music, as it is a mix of many elements.

Our first term we will be looking at is our intercept, which is, 2269562875.247, when all other variables are zero. This intercept is outside of the scope of our popularity, however, a song would not be able to have a zero value for all variables.

Acousticness - acousticness has a negative regression coefficient of -102.7, higher-order terms up to the power of 4, as well as interactions with: danceability, energy, explicit, key, tempo, year, liveness, and valence.

Energy - Energy has a positive regression coefficient of 3.206, higher-order terms up to the power of 3, as well as interactions with: instrumentals, key, speechiness, tempo, liveness, valence and mode.

Instrumentals - Instrumentals have a positive regression coefficient of 76.46, higher-order terms up to the power of 4, and an interaction with: key, tempo, year, liveness, and valence.

Speechiness - Speechiness has a positive regression coefficient of 123.05, along with a positive higher-order squared term, which means that it does apply a convex shape to our model. Speechiness also has an interaction with: year, key, energy, and explicit.

Tempo - Tempo has a positive regression coefficient of 0.216, higher-order terms up to the power of 5, and interacts with the following variables: acousticness, danceability, energy, explicit, instrumentals, year, liveness, and valence.

Year - We have included our Year term to be logged, higher-order, and as well be an important interaction term. With our log(year) term having a negative coefficient of 394351255.7, and our year term having a positive coefficient of 598873.5, year does have a large effect on the popularity of a song. Year also has an interaction with the following: acousticness, danceability, explicit,instrumentals, key, tempo, liveness, and mode

Liveness - Liveness has a positive regression coefficient of 51.217 and interactions with: acousticness, energy, instrumentals, and tempo.

Valence - valence has a negative regression coefficient of -0.6079 and interactions with: acousticness, danceability, energy, explicit, instrumentalness, key, tempo, liveness, and mode.

Explicit - explicit is a binary categorical variable with a negative regression coefficient of -270.8 when a song is explicit. Explicit also has interactions with: acousticness, danceability, instrumentalness, key, speechiness, tempo, year, valence, and mode.

Key - key is a categorical variable with 12 levels and interactions with: acousticness, danceability, energy, explicit, instrumentalness, speechiness, year, and valence.

Mode - mode has a positive regression coefficient of 22.2468 as well as interactions with: energy, explicit, year, and valence.

It isn't a surprise that every term has an interaction with three or more other terms. When we break down what each variable measures, other factors would definitely have an effect. If we take Liveness for example, it interacts with acousticness, energy, instrumentals, and tempo. During live performances, artists have a free range of how they would want to perform a different version of their music, which would have an effect on other terms during the song.

Our Model with Outliers Removed

When looking at our model we decided to also look at removing any leverage points that were above our 3*p/n threshold. After re-working our model, it looks like our R^2_{adj} dropped to 0.7739, and our RMSE increased to 10.29. As our data set is a data point that is entered automatically, we do not seem to think that there would be any clerical error in collecting this information. With the nature of how some music is, we thought it would be best to include all data points into our model. It does not change any other assumptions for our model.

```
Residual standard error: 10.29 on 165136 degrees of freedom
Multiple R-squared:  0.7741,    Adjusted R-squared:  0.7739
F-statistic:  3472 on 163 and 165136 DF,  p-value: < 0.00000000000000022
```

How well does our model predict popularity?

In order to assess our model's predictive potential, we used a technique from machine learning which involves splitting the data set into a training set and a testing set. The training set is a random sampling of 99% of the data, while the remaining 1% is reserved for testing the model's accuracy. In a departure from the usual rigor of a proper machine learning model, the same model we used above was used again but the regression coefficients were learned with only the reduced 'training data'.

```{r}
train_index = sample(1:nrow(spotify_data),0.99*nrow(spotify_data))
test_index = setdiff(1:nrow(spotify_data), train_index)

train_data = spotify_data[train_index,]
test_data = spotify_data[test_index, ]
#head(test_data) #1707 rows
```

	valence <dbl>	year <int>	acousticness <dbl>	danceability <dbl>	duration_ms <int>	energy <dbl>	explicit <int>
33	0.185	1921	0.505	0.233	686664	0.00817	0
171	0.414	1922	0.974	0.273	168107	0.31800	0
294	0.741	1923	0.937	0.625	176400	0.20300	0
423	0.871	1924	0.951	0.826	175720	0.44700	0
487	0.556	1924	0.995	0.334	168027	0.38200	0
535	0.918	1924	0.977	0.583	175200	0.30100	0

6 rows | 1-8 of 15 columns

```{r}
besttest = lm(popularity ~ acousticness + I(acousticness^2)+ I(acousticness^3)+ I(acousticness^4) +
danceability + energy + I(energy^2)+ I(energy^3) + factor(explicit) + instrumentalness+ I(instrumentalness^2)+
I(instrumentalness^3)+ I(instrumentalness^4) + factor(key) + speechiness+ I(speechiness^2)+ tempo +
I(tempo^2)+ I(tempo^3)+ I(tempo^4)+ I(tempo^5) + log(year)+I(year^2)+I(year^3) + liveness + valence  +
factor(mode) + acousticness*danceability + acousticness*energy + acousticness*factor(explicit) +
acousticness*factor(key) + acousticness*tempo + acousticness*year + acousticness*liveness  +
acousticness*valence + danceability*factor(explicit) + danceability*instrumentalness +
danceability*factor(key) + danceability*tempo + danceability*year  + danceability*valence +
energy*factor(explicit) + energy*instrumentalness + energy*factor(key) + energy*speechiness + energy*tempo +
energy*liveness + energy*valence + energy*factor(mode) + factor(explicit)*instrumentalness +
factor(explicit)*factor(key) + factor(explicit)*speechiness + factor(explicit)*tempo  + factor(explicit)*year
+ factor(explicit)*valence   + factor(explicit)*factor(mode) + instrumentalness*factor(key) +
instrumentalness*tempo + instrumentalness*year + instrumentalness*liveness + instrumentalness*valence +
factor(key)*speechiness + factor(key)*year + factor(key)*valence + speechiness*year + tempo*year +
tempo*liveness+ tempo*valence + year*liveness + year*factor(mode) + liveness*valence + liveness*factor(mode) +
valence*factor(mode), data = train_data) # this is the same bestfit model, only trained on a reduced data set


```
```{r}

test_results = data.frame(y = integer(), yhat = integer())

for(i in 1:nrow(test_data)){
  test_results[i, "yhat"] = predict(besttest, test_data[i,]) # store what the model predicts for popularity
  test_results[i, "y"] = test_data[i,"popularity"] # store the real popularity value
}


```
```{r}
SSE = 0
for(i in 1:nrow(test_results)){
  SSE = SSE + (test_results[i, "y"] - test_results[i,"yhat"])^2
}
MSE = SSE/nrow(test_results)
RMSE = sqrt(MSE)
RMSE
```

```
[1] 10.70669
```

	y	yhat
1	5	13.41672501
2	0	4.83713425
3	5	9.40689389
4	4	9.07817865
5	0	5.29474315
6	0	6.03538728
7	0	3.31733882
8	22	4.71337463
9	17	4.48550527
10	0	9.74518478
11	0	3.17386689
12	13	4.49623628
13	9	2.06739438
14	3	-0.83326115
15	0	7.18064357
16	0	-0.18221719
17	0	1.63066825
18	0	0.62892328
19	0	-0.05688119
20	0	-1.85227022
21	6	4.08211911
22	0	-2.34204860

The code snippets above shows the procedure for splitting and testing the data, as well as a subset of the test_results data frame. The test_results data frame stores the actual versus predicted values of popularity in the testing data. The RMSE of testing on 1% of the data came out to 10.70669. This is within range of the RMSE values for the other models we trained, but worse than our best fit model by 0.4467. Additionally, most predictions in the test results data frame are not close to the actual value for popularity.

Results and Discussion

After finding our final best fit model, we have decided to stick the course with our Data Science careers, and not to transition into the best and most popular musical act in the world. Trying to predict the popularity of a song using a regression model is a bit more challenging than it seems. Using the metrics we found in our data set, music is a more complex form and has many different factors when it comes to predicting the popularity of a song.

With having so many predicting terms in our model, we do also question if our model has an issue with overfitting. We do understand that there are many interaction terms,in addition to add higher-order terms within our model while they stay significant, however, with the introduction of many variables, we do have a concern about overfitting, which would inflate our R^2_{adj} value.

This is also in tune to an assumption which our model had failed, the homoscedasticity assumption. With our model not having equal variance error terms, our model may have provided us with prediction variables that are shown significant, however in reality they are not. This would also cause an overfitting of our model, and inflate our R^2_{adj} value, along with a lower RMSE.

As previously mentioned, we do also have a concern with outliers. It would seem that many of our entries would be considered an outlier depending on the method we have chosen to seek out these outliers. Knowing the variety of differences in music, we do not suspect these outliers to be clerical errors but just the nature of music.

We did also have some of our assumptions fail, including the Normality assumption. Initially, we tried to use the BoxCox transformation method, however, our response variable had non-positive variables. We figured it would not be ethical to remove any data points that do not work with our model. Then we tried to log our response variable, however, we ran into the same problem with the BoxCox transformation. The next step would be to log the independent variables, however, a lot of our independent variables included zeros as well and the only variable we were able to log was the year. We did try to correct our model for this, however, we believe that this would be past the scope of our current class.

In terms of improving the model, the obvious first step would be to include which artist released the song. Popular artists have many fans that will listen to every song they release while most artists on Spotify do not have nearly the same size as the audience. Our data set did include information about artists, but unfortunately, this information was a simple list of strings of artist names on the song listing. Including a list of artists in the regression model would not be a tenable solution as it would have made an already complicated model even more complex. Spotify also has an important variable about artists called ‘Monthly Listeners’. This variable would be an invaluable addition to our model of predicting song popularity. For example, Drake is a popular artist with ~53.6 million monthly listeners, while a young musician by the name of OMY has a mere 83 monthly listeners. We expect that the popularity of an artist would be tightly correlated with the popularity of their songs. Future iterations of our model would involve more data wrangling to include an artist’s ‘monthly listeners’ in the model.

References

Ay, Y. (2020, November 25). Spotify Dataset 1921-2020, 160k+ Tracks. Retrieved December 06, 2020, from https://www.kaggle.com/yamaerenay/spotify-dataset-19212020-160k-tracks

McNeese, B., Dr. (2011, June 19). Anderson-Darling Test for Normality. Retrieved December 06, 2020, from https://www.spcforexcel.com/knowledge/basic-statistics/anderson-darling-test-for-normality

Tutorial: Building train and test sets with the same characteristics. (2020, November 13). Retrieved December 06, 2020, from https://cran.r-project.org/web/packages/dataPreparation/vignettes/train_test_prep.html

Bronson, F. (2016, November 25). Hot 100 55th Anniversary: The All-Time Top 100 Songs. Retrieved December 07, 2020, from https://www.billboard.com/articles/list/2155531/the-hot-100-all-time-top-songs