What Makes Mainstream Music So Popular?

Pete Schultz

1 Introduction

What is it about mainstream music that makes it even more popular than other well known tracks? What specific features do music tracks share and how much influence do they have on popularity? I predict that time of release, danceability, language content, artist popularity, and duration are important factors to predicting a song's overall popularity. In this analysis, I will show through a regression model that these variables are indeed influential.

2 Data

The music data used for this analysis was provided by Spotify with access to their open API. I generated the dataset on May 15, 2020 by looping through each of Spotify's playlists and collecting all of the artists that were listed. From there I extracted song data from Spotify API of the top 10 songs from each artist. Altogether this filled my dataset with a total of 288,844 song records.

It is estimated that Spotify holds over 50 million tracks, so this dataset serves as a sample of the overall population. To that end, this sample is bias towards the popularity variable as it only includes the top 10 most popular songs of each artist collected. With the focus of using only the most popular music, I subset the dataset to only include songs with a popularity ranking between 70 and 90 (on a scale 1-100). This results in a dataset of 4121 song records to begin working with.

2.1 Variables

For this analysis I will be using observations of individual songs to discover whether the month of release, language content, danceability, artist popularity, and duration have an effect on its overall popularity.

2.2 Numerical Variables

Summary Statistics of Numerical Variables

Statistic	N	Mean	St. Dev.	Min	Pctl(25)	Pctl(75)	Max
Popularity	3,679	74.72	4.29	70	71	77	90
Danceability	3,679	0.65	0.16	0.00	0.60	0.80	1.00
Artist Followers	3,679	99.65	144.48	0.001	15.22	114.15	1,264.52

2.2.1 Popularity

Popularity is placed into values between 0 and 100, with 100 being the most popular. Spotify calculates popularity "by algorithm and is based, in the most part, on the total number of plays the track has had and how recent those plays are."[1] Because Spotify uses recent plays to boost popularity, I restricted the dataset to a maximum popularity ranking of 90 as most songs over the latter are fairly new.

2.2.2 Danceability

"Danceability describes how suitable a track is for dancing based on a combination of musical elements including tempo, rhythm stability, beat strength, and overall regularity. A value of 0.0 is least danceable and 1.0 is most danceable." For simplicity purposes, I rounded each value to the first digit.

2.2.3 Artist Followers

Spotify provides fans of particular artists the option to follow them in the app. Doing so allows users to receive notifications whenever the artist releases new music. This keeps fans updated and return helps the artist. For ease of interpretation, I divided artist followers by a factor of 50,000.

---

[1] https://developer.spotify.com/documentation/web-api/reference/#endpoint-get-an-albums-tracks

Univariate Statistics for Categorical Variables		
Category	**Frequency**	**Percent**
Content		
Clean	2557	69.5%
Explicit	1122	30.5%
Duration		
< 2:00	43	1.2%
2:00-2:30	191	5.2%
2:30-3:00	617	16.8%
3:00-3:30	1075	29.2%
3:30-4:00	916	24.9%
> 4:00	837	22.8%
Month		
January	411	11.2%
February	281	7.6%
March	378	10.3%
April	389	10.6%
May	317	8.6%
June	265	7.2%
July	203	5.5%
August	264	7.2%
September	258	7%
October	323	8.8%
November	372	10.1%
December	218	5.9%

2.3 Categorical Variables

2.3.1 Content

Songs listed on Spotify that contain any such use of profanity are marked *explicit*.

2.3.2 Duration

In Spotify's API, the track length is provided in milliseconds. From this I placed each track to fall into a specific duration category (units = minutes). These categories are listed: < 2:00, 2:00-2:30, 2:30-3:00, 3:00-3:30, 3:30-4:00, and > 4:00.

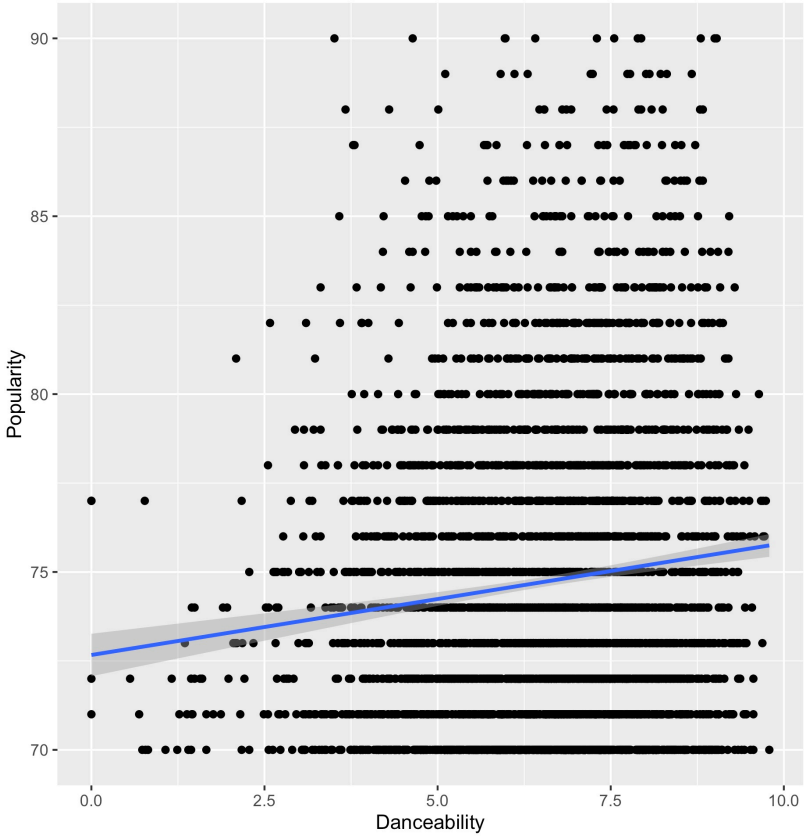
2.3.3 Month

The release date for each of the songs in the dataset are formatted in 'YYY-MM-DD' to which I broke into a factor by months.

3 Results and Interpretation

3.1 Model 1

3.1.1 Pre-model Inspection



3.1.2 Model 1 Summary

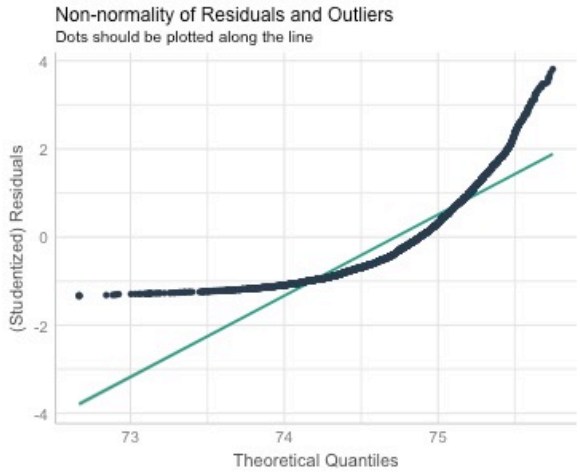
Model 1: Characteristics associated with Song Popularity

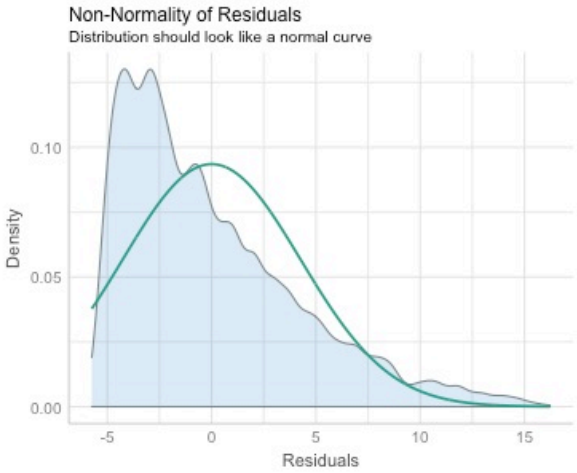
Predictor	Estimate	Std. Error	t-statistic	p-value
Intercept	72.666	0.304	238.97	< 0.001 ***
Danceability	0.315	0.045	6.95	< 0.001 ***

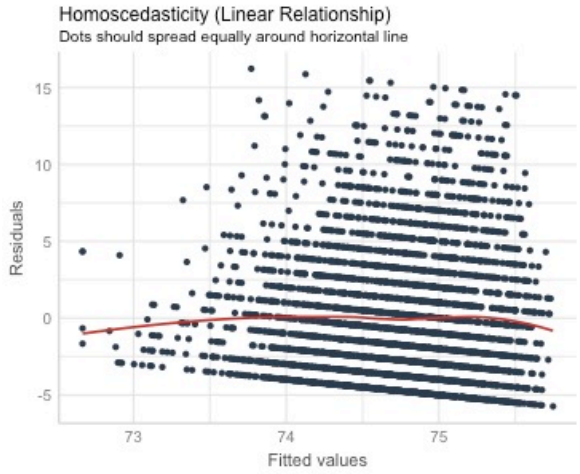
Note:
n = 3679. r-squared = 0.01, F(1,3677) = 48.34.

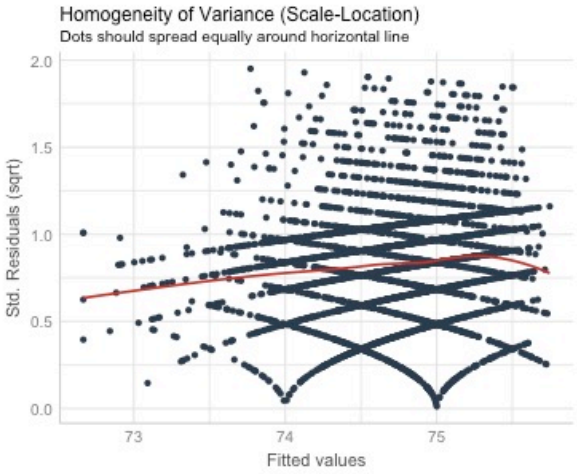
The coefficient for danceability is 3.15 (not 0.315) because for every .1 increase (from 0 to 1.0) in danceability, popularity grows an extra 3.15 units. The p-value is lower than an alpha of 0.05, which means that danceability is a significant predictor of popularity. The r-squared is 0.01 which means that danceability explains 1% of the variance in popularity. The p-value for the F-test is below 0.05, which means that including danceability into the model significantly outweighs that of a null model.

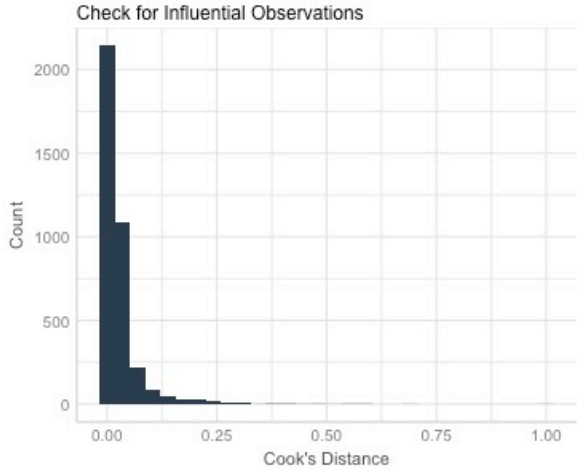
3.1.3 Model 1 Assumptions Check





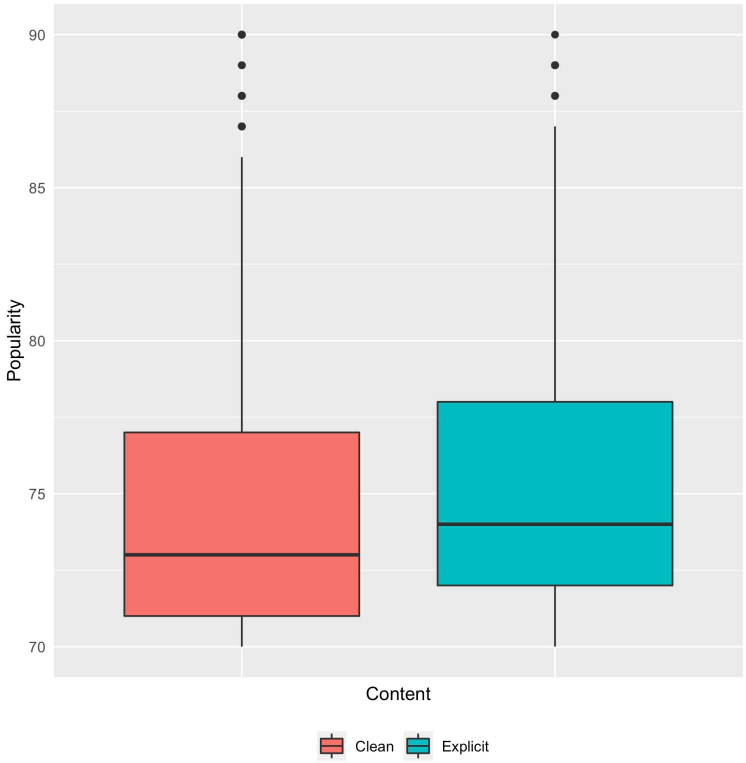






3.2 Model 2

3.2.1 Pre-model Inspection



3.2.2 Model 2 Summary

Model 2: Characteristics associated with Song Popularity

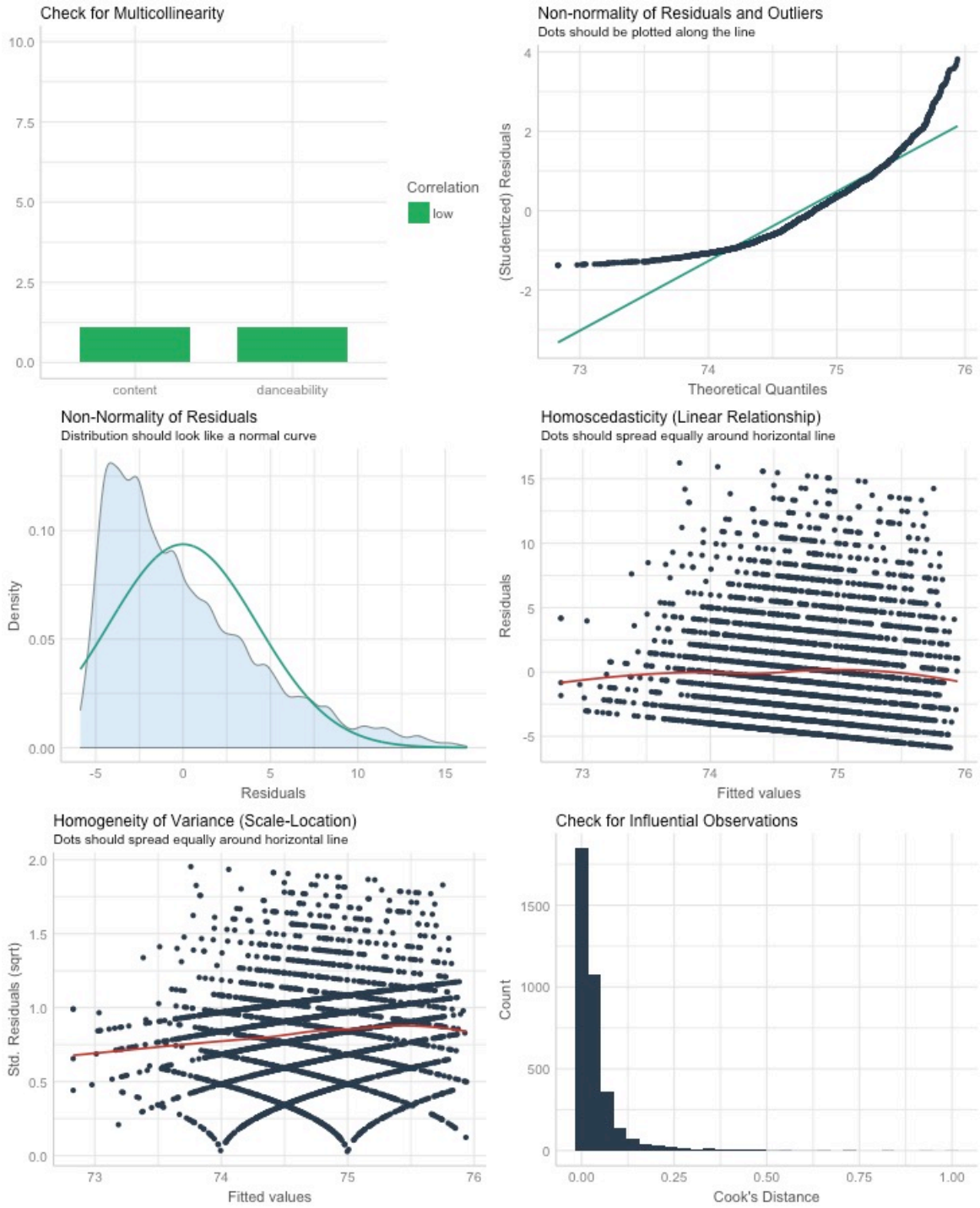
Predictor	Estimate	Std. Error	t-statistic	p-value
Intercept	72.829	0.308	236.74	< 0.001 ***
Danceability	0.265	0.048	5.57	< 0.001 ***
Explicit	0.530	0.161	3.30	< 0.001 ***

Note:
n = 3679. r-squared = 0.02, F(2,3676) = 29.67.

Holding danceability constant, the coefficient for Language Content (Explicit) is 0.5, meaning on average explicit songs are 0.5 more popular than clean songs. The p-value is lower than an alpha of 0.05, which means that language content is a significant predictor of

popularity. The r-squared is 0.02 which means that language content explains 2% of the variance in popularity.

3.2.3 Model 2 Assumptions Check



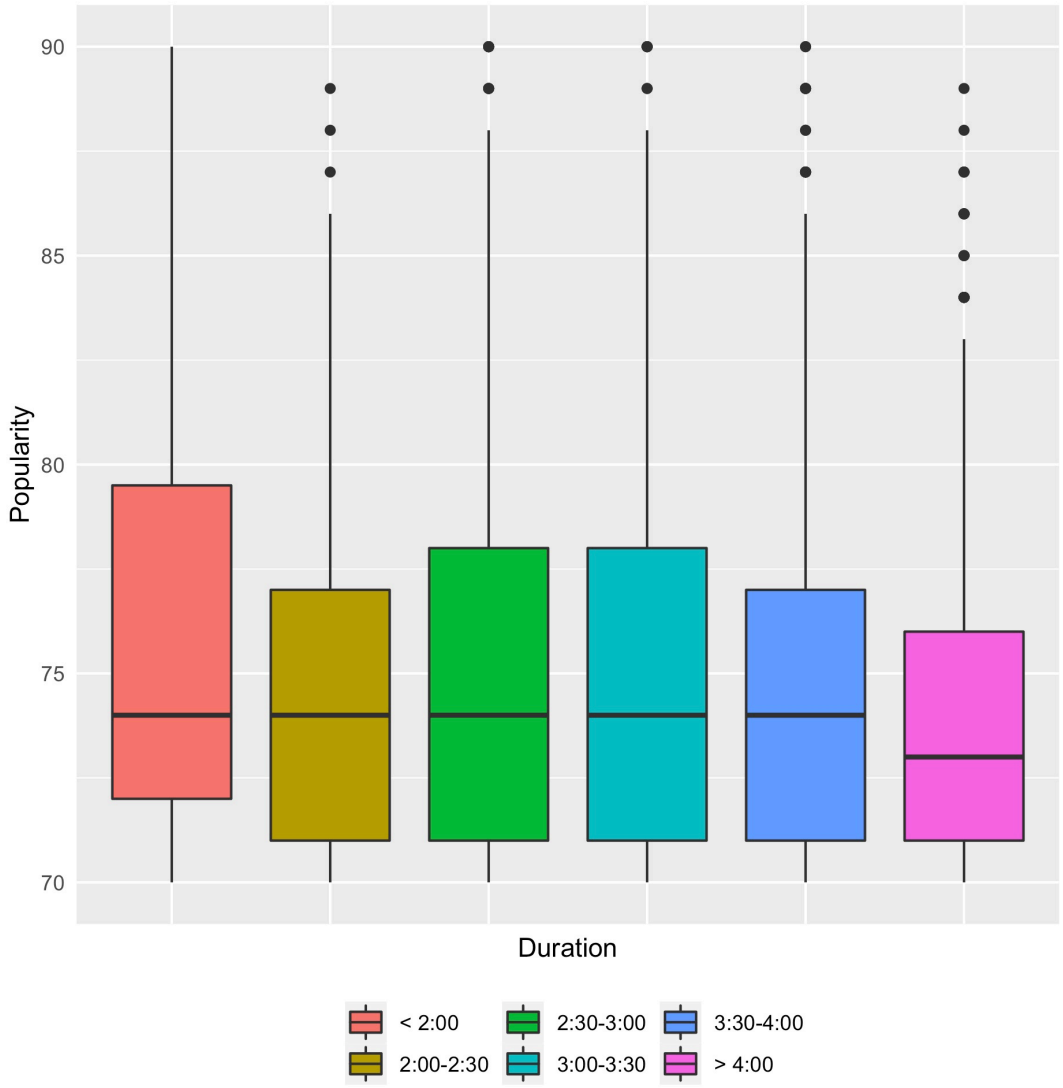
3.2.4 Model Comparison

Comparison: Model 1 vs. Model 2					
Residuals Degrees of Freedom	**Residual Sum of Squares**	**Degrees of Freedom**	**Sum of Squares**	**T Statistic**	**p-value**
3676	66739.15	1	197.1499	10.85904	< 0.001 ***

The p-value is below 0.05, which means that the new model has a significant difference from the old model.

3.3 Model 3

3.3.1 Pre-model Inspection



3.3.2 Model 3 Summary

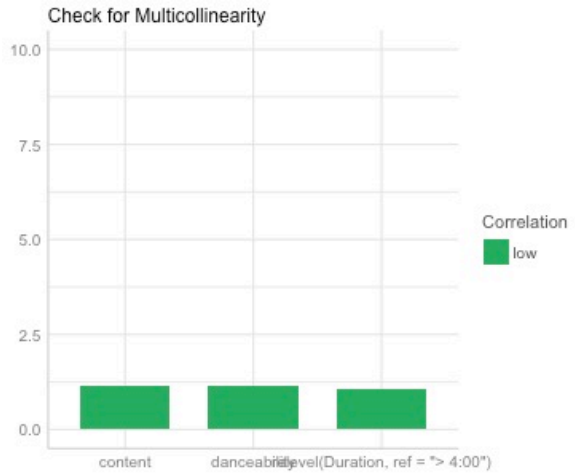
Model 3: Characteristics associated with Song Popularity

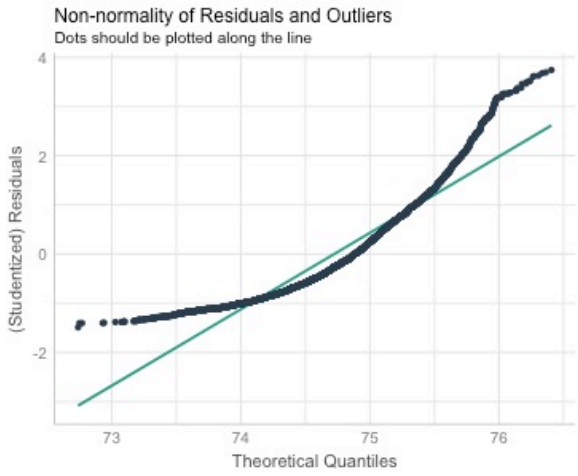
Predictor	Estimate	Std. Error	t-statistic	p-value
Intercept	72.574	0.319	227.72	< 0.001 ***
Danceability	0.238	0.048	4.94	< 0.001 ***
Explicit	0.532	0.162	3.28	0.001 **
< 2:00	1.047	0.669	1.57	0.118
2:00-2:30	0.366	0.346	1.06	0.290
2:30-3:00	0.619	0.229	2.71	0.007 **
3:00-3:30	0.720	0.198	3.64	< 0.001 ***
3:30-4:00	0.332	0.205	1.62	0.105

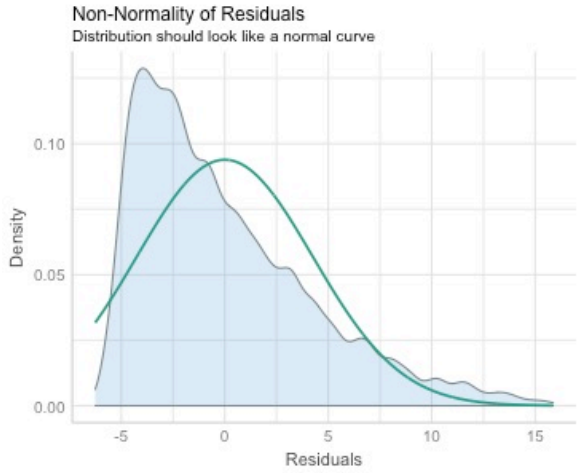
Note:
n = 3679. r-squared = 0.02, F(7,3671) = 10.76.

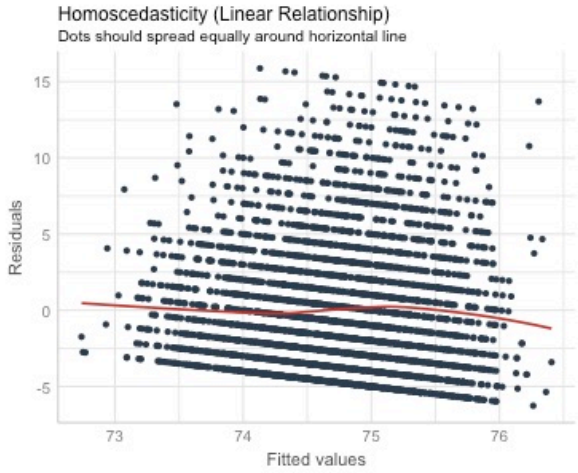
Holding danceability and language content constant, the coefficients for duration are quite interesting. Songs with duration < 2 minutes are 1.047 more popular, songs with duration 2-2:30 minutes are 0.366 more popular, songs with duration 2:30-3 minutes are 0.619 more popular, songs with duration 3-3:30 minutes are 0.72 more popular, and songs with duration 3:30-4 minutes are 0.332 more popular than other songs. The p-value is lower than an alpha of 0.05 for duration 2:30-3 minutes and 3-30 minutes, which means that these durations are significant predictors of popularity. The r-squared is 0.02 which means that song duration explains 2% of the variance in popularity.

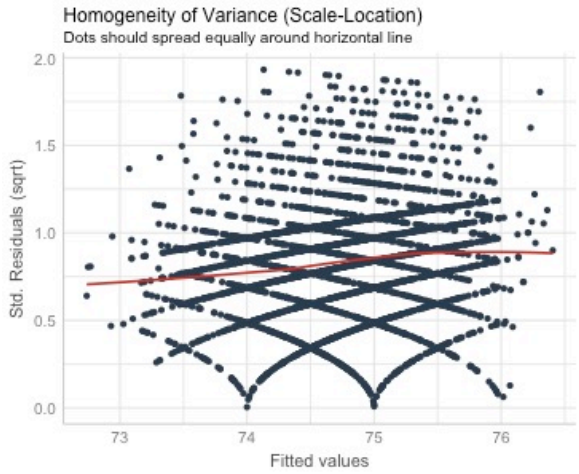
3.3.3 Model 3 Assumptions Check

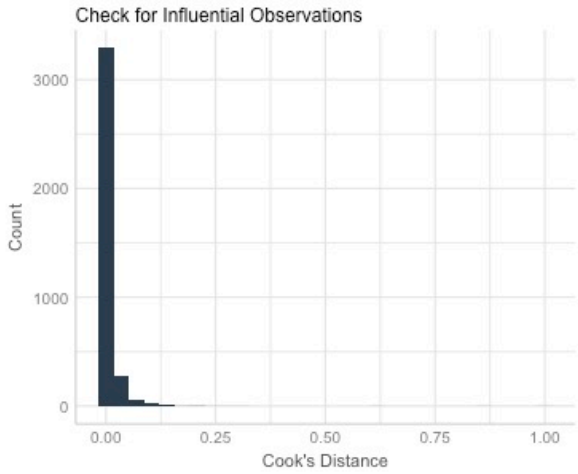












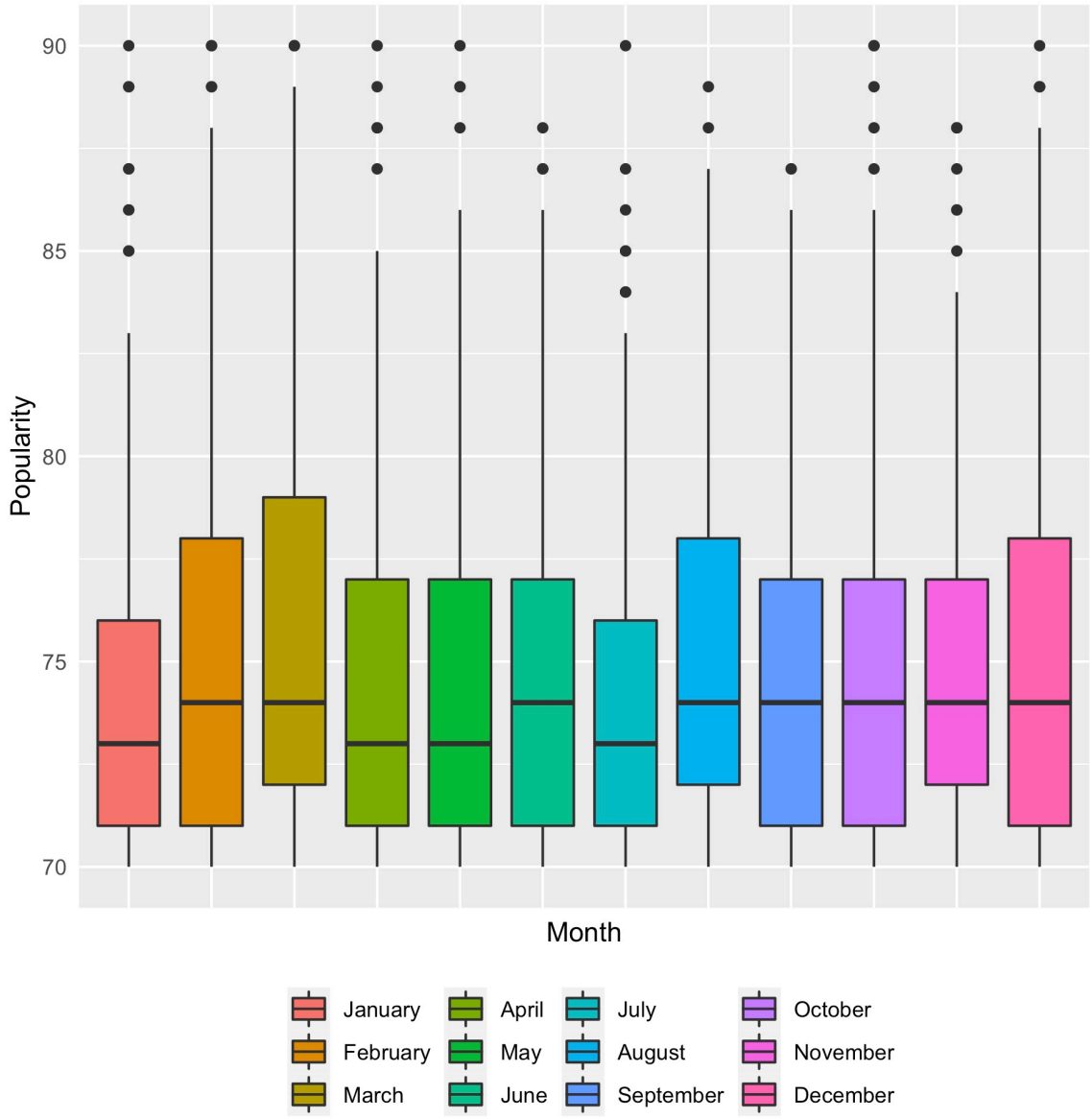
3.3.4 Model Comparison

Comparison: Model 2 vs. Model 3					
Residuals Degrees of Freedom	**Residual Sum of Squares**	**Degrees of Freedom**	**Sum of Squares**	**T Statistic**	**p-value**
3671	66452.34	5	286.8083	3.168807	0.007 **

The p-value is below 0.05, which means that the new model has a significant difference from the old model.

3.4 Model 4

3.4.1 Pre-model Inspection



3.4.2 Model 4 Summary

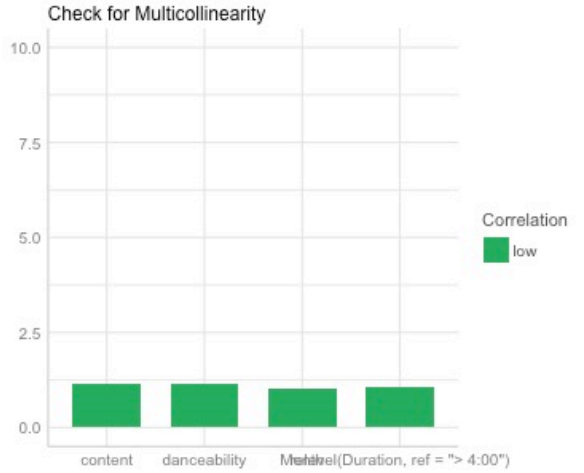
Model 4: Characteristics associated with Song Popularity

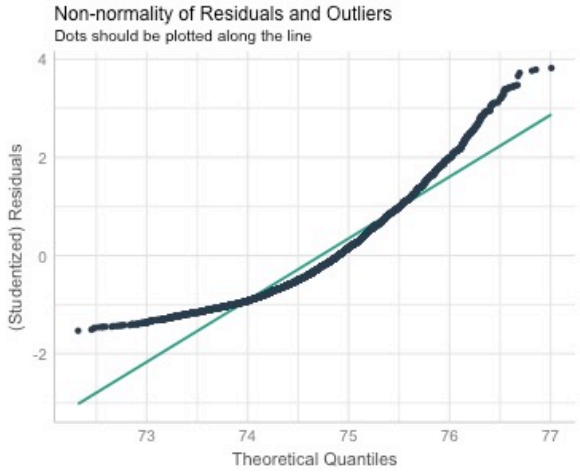
Predictor	Estimate	Std. Error	t-statistic	p-value
Intercept	71.962	0.374	192.48	< 0.001 ***
Danceability	0.240	0.048	4.98	< 0.001 ***
Explicit	0.489	0.162	3.01	0.003 **
< 2:00	0.914	0.667	1.37	0.171
2:00-2:30	0.341	0.345	0.99	0.324
2:30-3:00	0.564	0.228	2.47	0.014 *
3:00-3:30	0.699	0.197	3.54	< 0.001 ***
3:30-4:00	0.321	0.204	1.57	0.116
February	0.977	0.329	2.97	0.003 **
March	1.420	0.303	4.69	< 0.001 ***
April	0.384	0.301	1.28	0.202
May	0.309	0.317	0.97	0.330
June	0.881	0.334	2.64	0.008 **
July	-0.040	0.364	-0.11	0.912
August	1.180	0.335	3.52	< 0.001 ***
September	0.583	0.337	1.73	0.084 .
October	0.350	0.316	1.11	0.268
November	0.583	0.304	1.92	0.055 .
December	1.231	0.356	3.46	< 0.001 ***

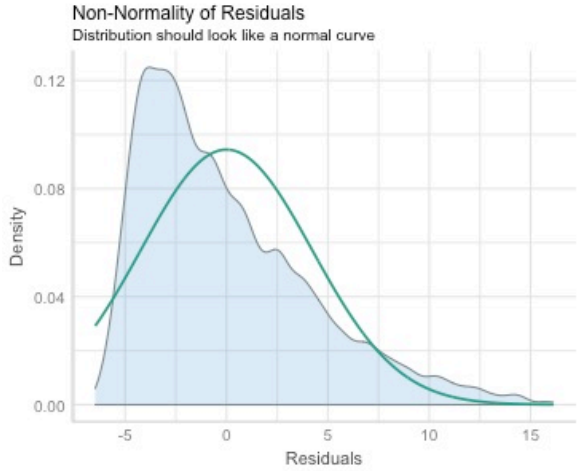
Note:
n = 3679. r-squared = 0.03, F(18,3660) = 6.64.

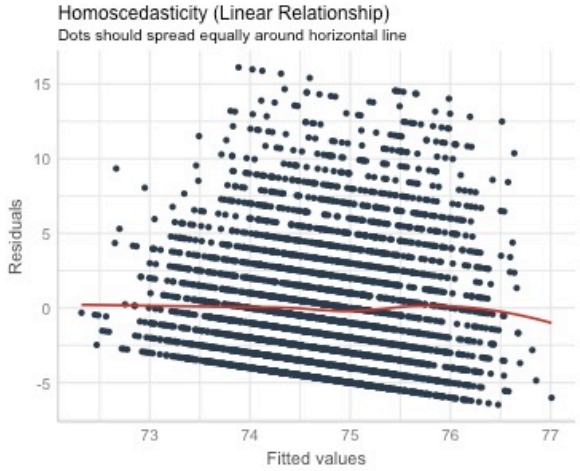
Holding danceability, language content, and duration constant, the coefficients for month of release are quite interesting. February, March, June, August, and December are the only months to consider as their p-value is < 0.05. February is 0.977 more popular, March is 1.42 more popular, August is 1.18 more popular, and December is 1.231 more popular than other songs.

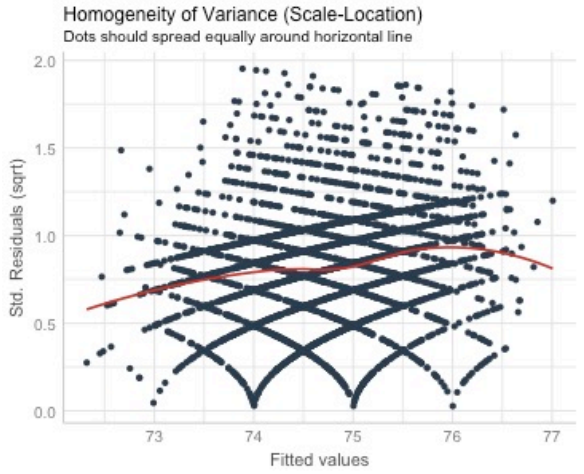
3.4.3 Model 4 Assumptions Check

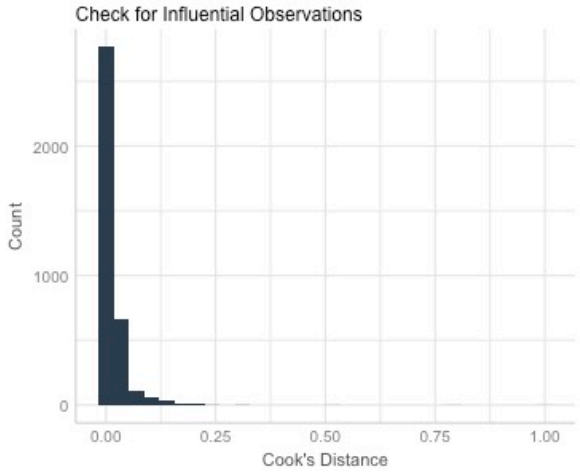












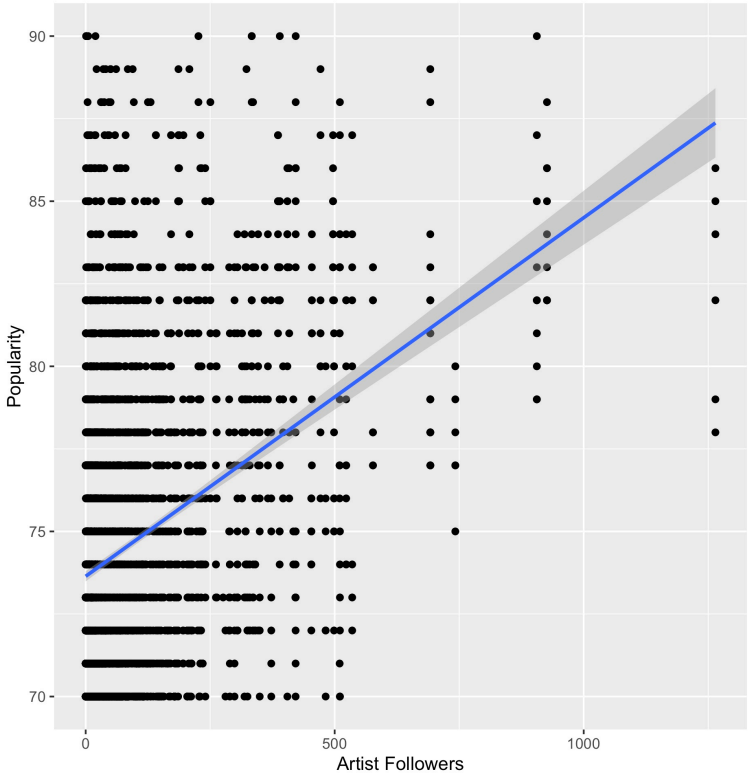
3.4.4 Model Comparison

Comparison: Model 3 vs. Model 4					
Residuals Degrees of Freedom	**Residual Sum of Squares**	**Degrees of Freedom**	**Sum of Squares**	**T Statistic**	**p-value**
3660	65673.27	11	779.0712	3.947089	< 0.001 ***

The p-value is below 0.05, which means that the new model has a significant difference from the old model.

3.5 Model 5

3.5.1 Pre-model Inspection



3.5.2 Model 5 Summary

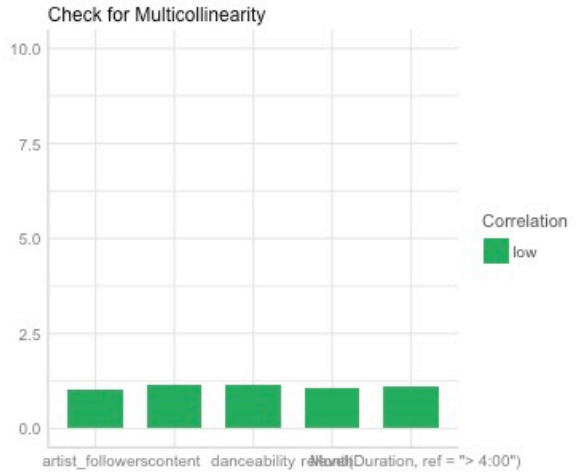
Model 5: Characteristics associated with Song Popularity

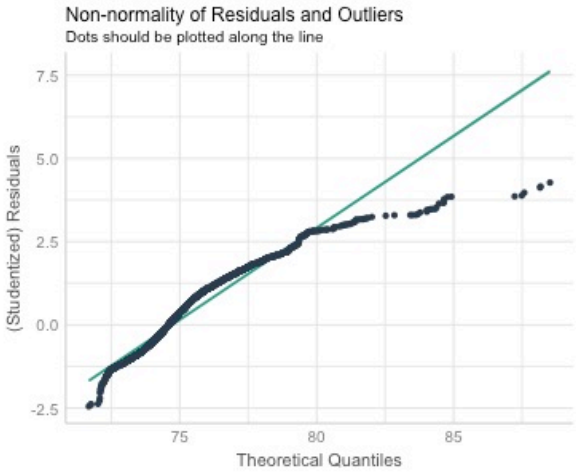
Predictor	Estimate	Std. Error	t-statistic	p-value
Intercept	71.271	0.348	204.59	< 0.001 ***
Danceability	0.189	0.045	4.21	< 0.001 ***
Explicit	0.494	0.151	3.28	0.001 **
< 2:00	1.573	0.620	2.54	0.011 *
2:00-2:30	0.918	0.321	2.86	0.004 **
2:30-3:00	0.982	0.213	4.62	< 0.001 ***
3:00-3:30	0.813	0.183	4.43	< 0.001 ***
3:30-4:00	0.260	0.190	1.37	0.171
February	0.730	0.305	2.39	0.017 *
March	1.166	0.282	4.14	< 0.001 ***
April	0.408	0.280	1.46	0.145
May	-0.110	0.295	-0.37	0.709
June	0.430	0.311	1.38	0.166
July	-0.042	0.338	-0.12	0.902
August	0.777	0.312	2.49	0.013 *
September	0.316	0.313	1.01	0.314
October	0.290	0.293	0.99	0.323
November	0.346	0.282	1.23	0.219
December	1.050	0.331	3.17	0.002 **
Artist Followers	0.011	0.000	24.19	< 0.001 ***

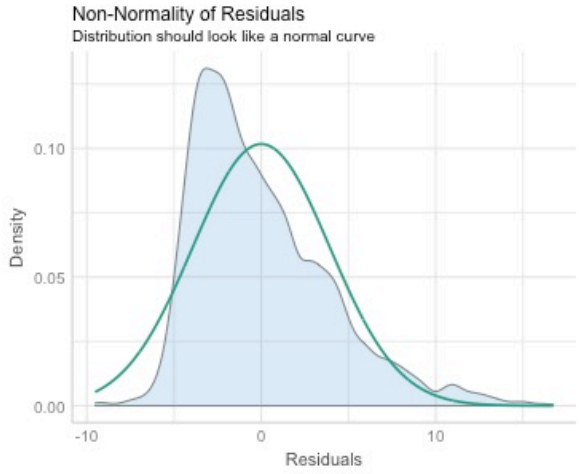
Note:
n = 3679. r-squared = 0.16, F(19,3659) = 38.09.

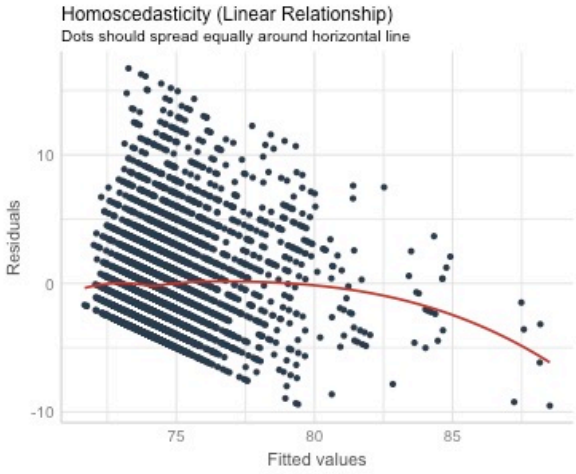
Holding danceability, language content, duration, and month of release constant, artist popularity is deemed significant to song popularity as its p-value is less than 0.05. Since artist followers is divided by 50,000 for ease of interpretation, this coefficient means that for every increase in 50,000 followers an artist has their songs increase by 0.011 in popularity.

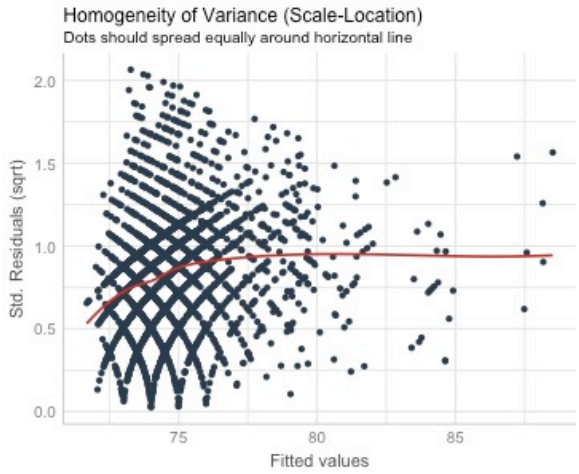
3.5.3 Model 5 Assumptions Check

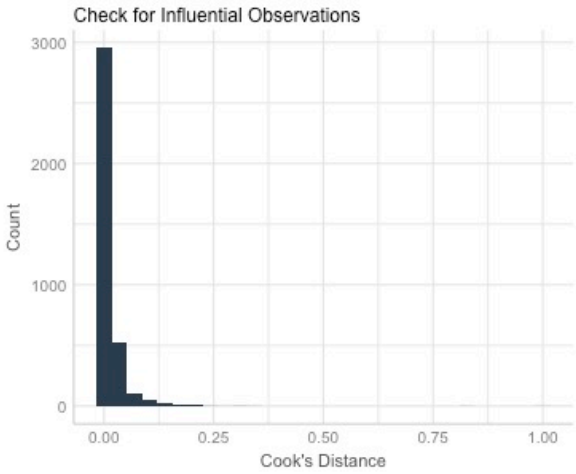












3.5.4 Model Comparison

Comparison: Model 4 vs. Model 5					
Residuals Degrees of Freedom	**Residual Sum of Squares**	**Degrees of Freedom**	**Sum of Squares**	**T Statistic**	**p-value**
3659	56617.31	1	9055.964	585.2587	< 0.001 ***

The p-value is below 0.05, which means that the new model has a significant difference from the old model.

4 Conclusion

The regression model resulted in some interesting findings relating to the features of time of release, danceability, language content, artist popularity, and duration. I found that mainstream music typically has the upper edge when it follows the following criteria: has a high danceability, has explicit language content, is either less than 2 minutes, between 2-2:30, between 2:30-3, or between 3-3:30, is released in February, March, August, or December, and has a high amount of followers on the music service (in this case Spotify).