Influential Predictors of Song Popularity

MSDS 601 Final Group 4
Tatshini Ganesan, Zoe Le, Chris Nishimura

I. Abstract

Problem

Our goal is to find the best model to predict the popularity of a track based on other metrics measured by Spotify and to gain inferential insight into which predictors are significant in this prediction. The popularity of a track is calculated based on the total number of plays the track has had and how recent those plays are. Since our data set is based on subjective traits that are difficult to measure, our R^2 is relatively low. Looking at other notebooks, the R^2 for baseline linear models was consistently below 0.1.

Methods

For our model, we chose to use multiple linear regression (MLR). Although MLR has lower predictive power than other models, such as random forest and neural networks, we want to maintain our ability to make inferences about predictors' significance.

In our project, we fit two MLR models: one on the full dataset and another on only songs under the K-pop genre. We use the following steps to validate our model

1. Check for multicollinearity with VIF scores
2. Fit the initial model
3. Remove influential points with external studentized residual and Cook's distance
4. Check heteroscedasticity with residual plot and Breusch-Pagan test
5. Check normality with QQ plot and Jacque-Bera test
6. Check non-linearity between x and y with scatter plot
7. Fit second model
8. Model selection with best subset method using adjusted R-squared and Mallows' Cp
9. Final model

II. About the Dataset

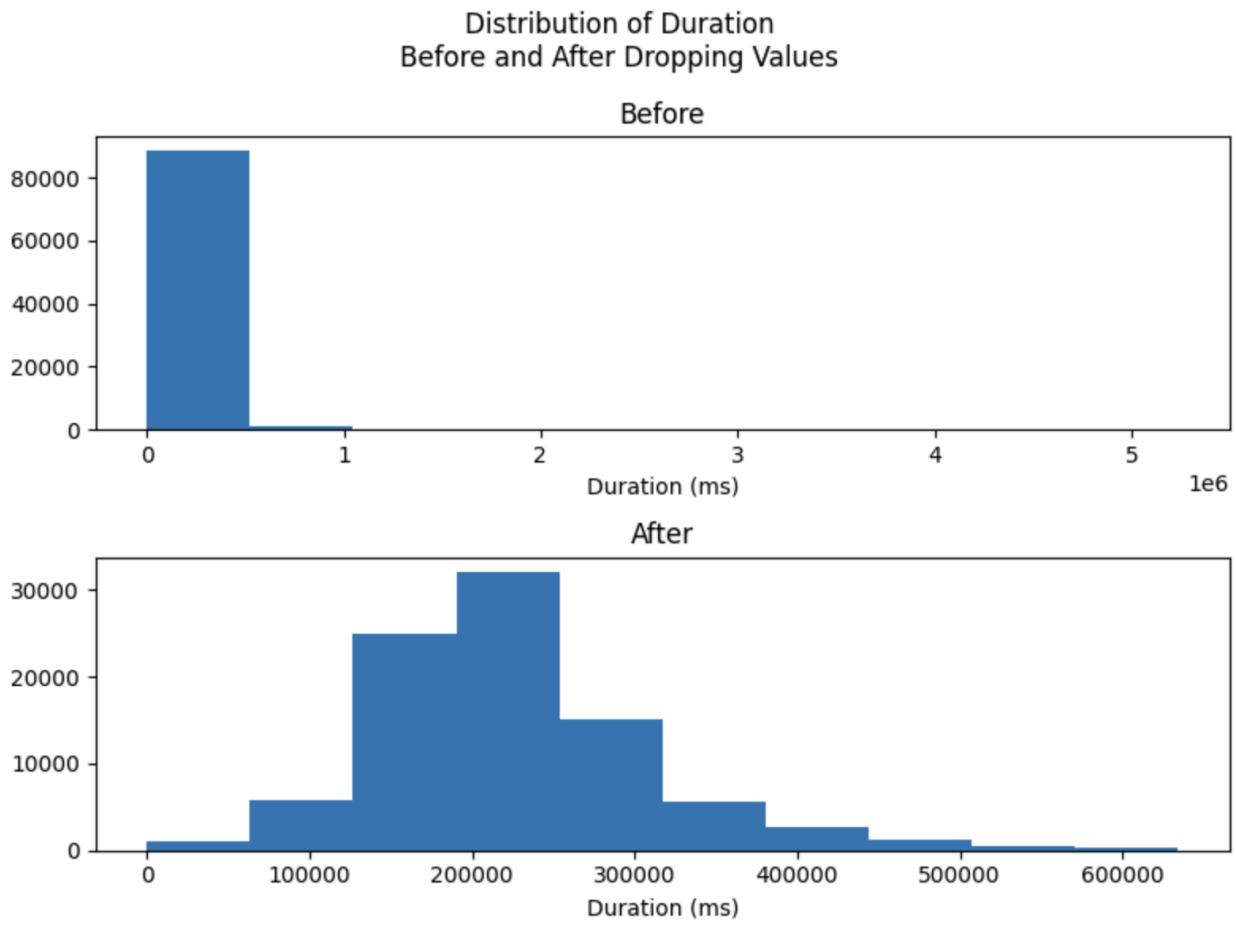
We chose to analyze the 'Spotify Tracks Dataset' from Kaggle. The uncleaned dataset has ~114,000 records and 21 features, including a column of row index.

Feature	Data type	Notes
track_id	string	Unique song identifier
artists	string	Artist name(s)

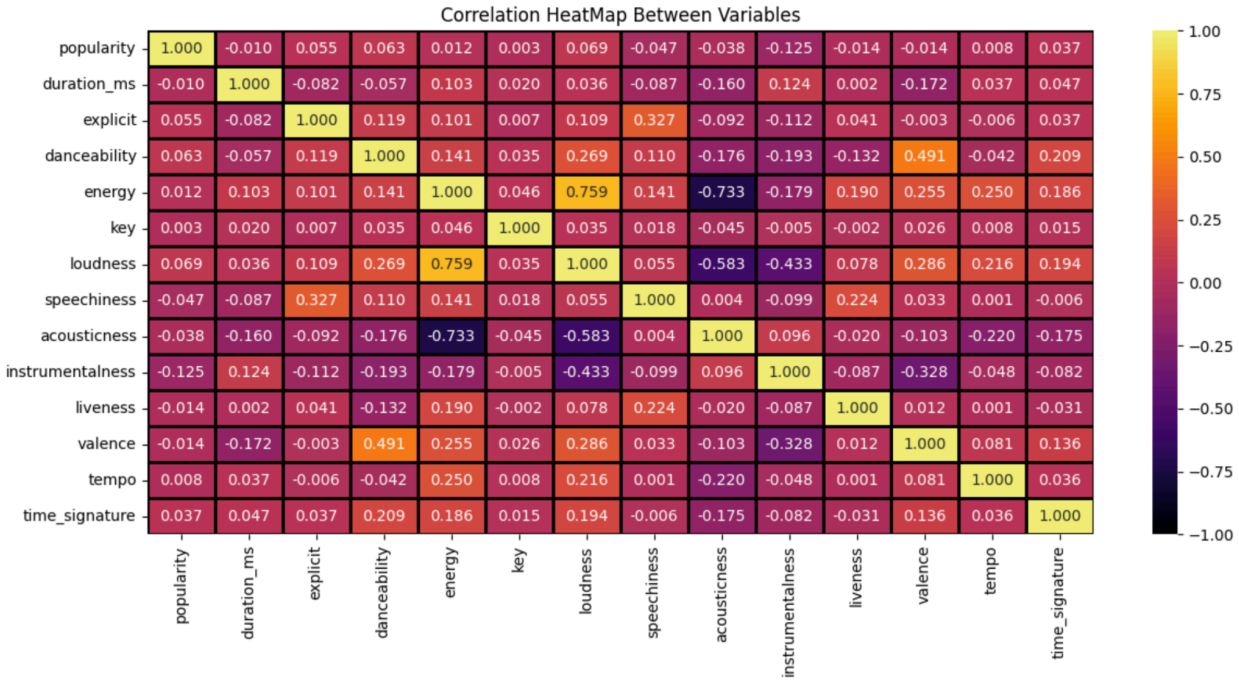
album_name	string	Album name
track_name	string	Name of song
popularity	int [0, 100]	Based on number of recent plays
duration_ms	int	Length of song in milliseconds
explicit	boolean string	Null value denoted with 'Unknown'
danceability	float [0, 1]	Based on tempo, rhythm stability, beat strength, and overall regularity
energy	float [0, 1]	Measures perceptual intensity and activity (i.e fast, loud, noisy)
key	categorical	Denotes the key of the song 0=C, 1=C#/Db, 2=D, …, -1=Null
loudness	float	Decibels
mode	binary	1 if major, 0 if minor
speechiness	float [0, 1]	0-0.33 indicates low likelihood of speech present. 0.33-0.66 indicates song with lyrics. 0.66-1 indicates spoken word mediums (i.e. podcast)
acousticness	float [0, 1]	Confidence measure of whether track is acoustic or not
instrumentalness	float [0, 1]	Predicts whether a track contains no vocals (1 is no vocals)
liveness	float [0, 1]	Detects the presence of an audience in the recording (close to 1, audience is likely present)
valence	float [0, 1]	Describes the musical 'positiveness' conveyed.
tempo	float	Estimated tempo of the track in BPM
time_signature	int	Estimated time signature (ranges from 3 to 7 but represents $\frac{3}{4}$ to $\frac{7}{4}$)
track_genre	string	Genre of the track. Songs get listed multiple times if listed in different genres

III. Explanatory analysis

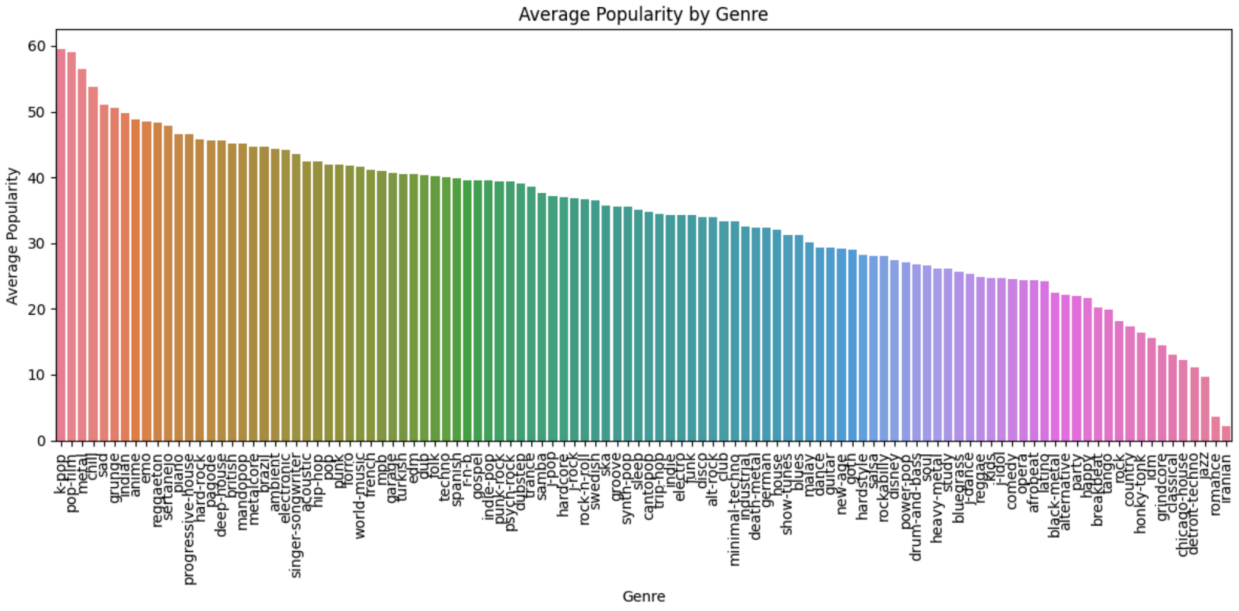
To find the influential musical factors that contribute to the popularity, we explore the correlation between each factor and popularity.



Looking at the initial distribution of song duration, we decided to remove outliers in order to reduce the skew of the distribution. After removing the top 0.5% of longest tracks, we observe a greatly improved, although still slightly right-skewed distribution for the duration among all songs. This indicates that there are more short songs than long songs.



From this heatmap, it can be observed that loudness and danceability show the highest correlations with popularity at 0.069 and 0.063 respectively. The lowest correlation, which is -0.125, pertains to instrumentalness. However, these correlations are quite low, indicating that these factors alone do not significantly impact the popularity of a song.



After removing duplicate genres, there are 110 unique genres. The bar plot displaying the average popularity by genre illustrates significant differences among them.
Among these 110 genres, K-pop, pop-film, metal, and chill genres exhibit the highest average popularity, all exceeding 50. In contrast, romance and Iranian genres have the lowest popularity, each registering at less than 5.

IV. Regression analysis

A, Feature selection

We first picked all predictors of interest including: duration_ms, explicit, danceability, energy, key, loudness, mode, speechiness, acousticness, instrumentalness, liveness, valence, tempo, time_signature, track_genre.

Features	VIF Factor	Multicollinearity level
Intercept	804.504040	
explicit[T.True]	1.294762	low
time_signature[T.1]	6.9024645283	moderate
time_signature[T.3]	50.573669024	high
time_signature[T.4]	65.434303766	high
time_signature[T.5]	12.034029068	high
...	...	...
energy	5.15724576	moderate
loudness	4.04396054	moderate
acousticness	3.096871	low
instrumentalness	2.159018	low
liveness	1.285332	low
valence	2.040537	low
tempo	1.162565	low

Based on the VIF values, we can conclude that time signature is highly impacted by multicollinearity. Since loudness is used to calculate energy, we can drop loudness too.
After dropping the time signature and loudness feature, we fit the first model: popularity~duration_ms+explicit+danceability+energy+key+mode+speechiness+acousticness+instrumentalness+liveness+valence+tempo+track_genre.

B, Model Diagnosis

Heteroscedasticity exists so we will use robust standard error instead to perform t tests. From the JB test and qq plot, we can see that the model violates normality assumption. However, the dataset is large and the violation is not severe and we can continue with model selection.

Model with track_genre

R-squared:	0.310
Adj. R-squared:	0.309
F-statistic:	307.4
Prob (F-statistic):	0.00
Log-Likelihood:	-3.7991e+05
AIC:	7.601e+05
BIC:	7.613e+05

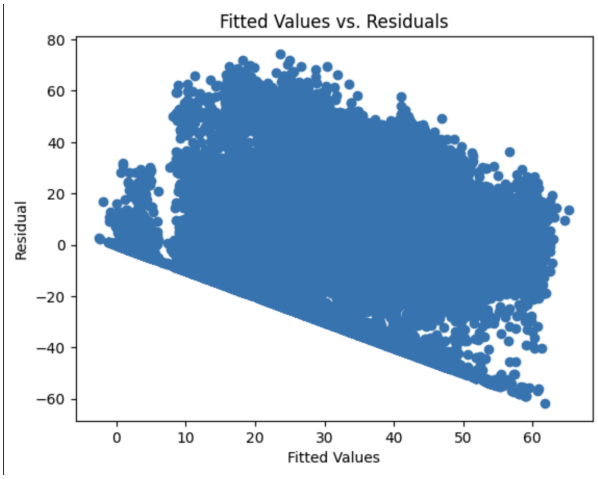
Omnibus:	1863.172	**Durbin-Watson:**	0.800
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	4088.381
Skew:	0.046	**Prob(JB):**	0.00
Kurtosis:	4.045	**Cond. No.**	2.42e+07

Model without track_genre

R-squared:	0.033
Adj. R-squared:	0.033
F-statistic:	132.5
Prob (F-statistic):	0.00
Log-Likelihood:	-3.9495e+05
AIC:	7.899e+05
BIC:	7.902e+05

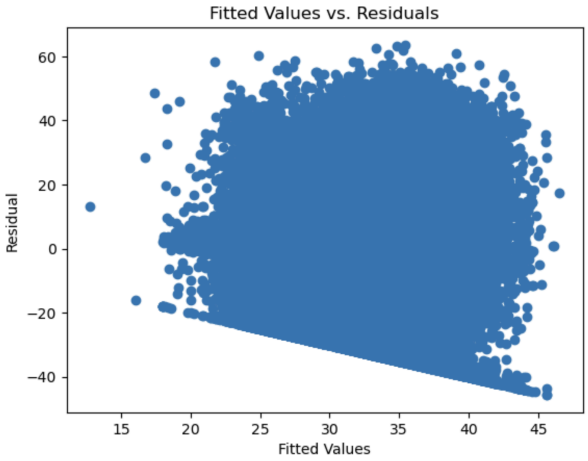
Omnibus:	3965.393	**Durbin-Watson:**	0.599
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	1641.672
Skew:	-0.015	**Prob(JB):**	0.00
Kurtosis:	2.336	**Cond. No.**	3.30e+06

BP test p-value: 0.00

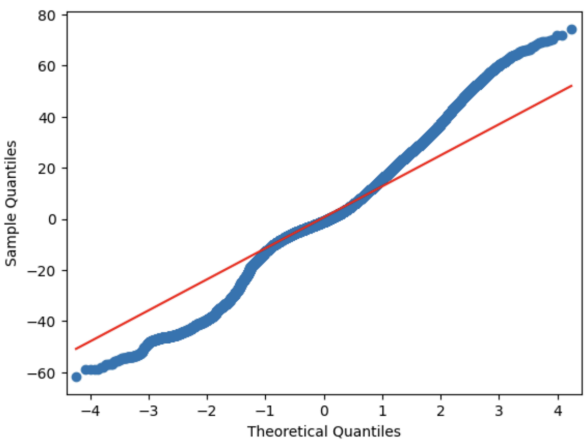


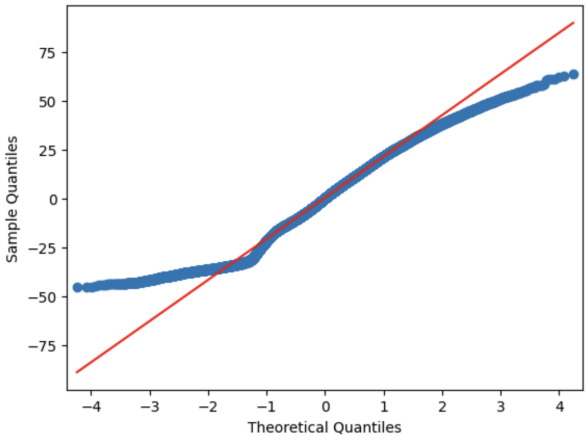
showed very little improvement regardless of log or boxcox transformation

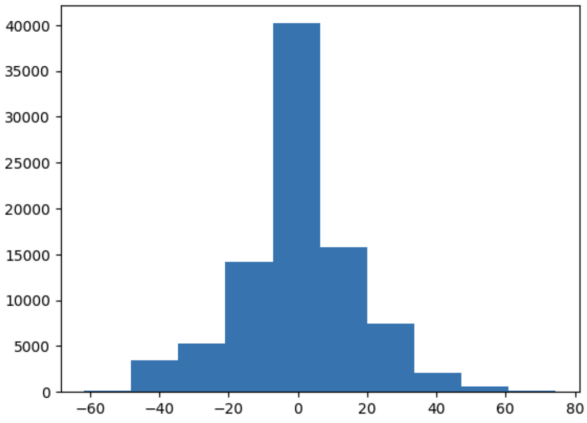
BP test p-value: 0.0

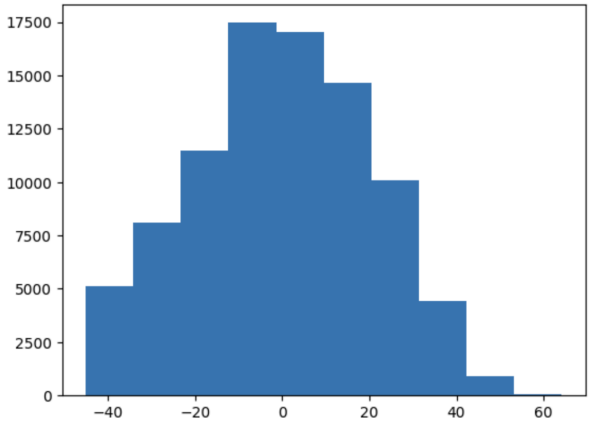


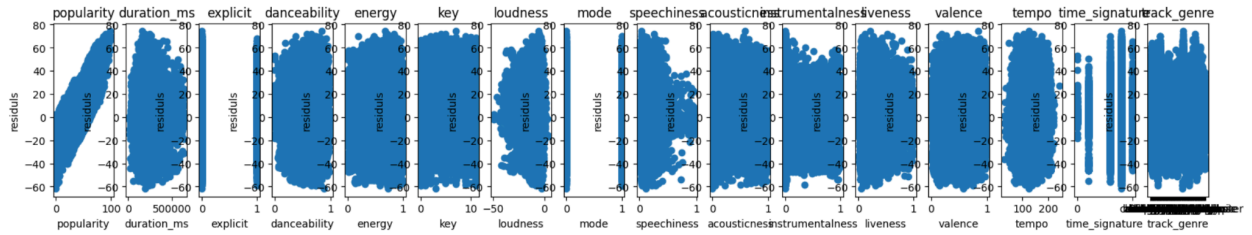
Similarly showed very little improvement regardless of transformation











From the scatter plot between each feature, it looks like residuals are proportional to speechiness so we decided to drop speechiness. After we drop these 2 features, test results don't change much.

C, Model Selection

For the model selection process, we chose to use the best subset method using Mallow's Cp, AIC and BIC, along with the adjusted R-squared. There are a few models with adjusted R2 ranging from 0.309270 to 0.309280, the best that we have got. We also narrowed down our model pool using Mallow's Cp to 3 models.

1. popularity ~ explicit + danceability + energy + key + mode + speechiness + acousticness + instrumentalness + liveness + valence + tempo + track_genre

2. popularity ~ explicit + danceability + energy + mode + speechiness + acousticness + instrumentalness + liveness + valence + tempo + track_genre
3. popularity ~ duration_ms + explicit + danceability + energy + mode + speechiness + acousticness + instrumentalness + liveness + valence + tempo + track_genre

We can clearly identify that the second model from the above list of models is the better compared to all other models using the AIC and BIC values

	Predictors	adj-R2	Cp	Num_pred	aic	bic
3	explicit+danceability+energy+loudness+mode+spe...	0.309280	110.294774	120.0	760651.126880	761788.485254
2	explicit+danceability+energy+key+loudness+mode...	0.309277	111.568385	121.0	760652.399416	761799.157445
4	duration_ms+explicit+danceability+energy+loudn...	0.309274	112.030674	121.0	760652.862390	761799.620420
5	explicit+danceability+energy+loudness+mode+spe...	0.309251	112.965936	119.0	760653.804821	761781.763538
0	duration_ms+explicit+danceability+energy+key+l...	0.309272	113.296517	122.0	760654.127144	761810.284829

Models sorted by Cp ascending

	sum_sq	df	F	PR(>F)
explicit	3.904298e+04	1.0	133.334548	8.035072e-31
track_genre	1.050332e+07	109.0	329.078849	0.000000e+00
danceability	3.015346e+04	1.0	102.976197	3.496280e-24
energy	6.738148e+03	1.0	23.011253	1.613156e-06
loudness	2.257454e+03	1.0	7.709366	5.494649e-03
mode	1.312536e+04	1.0	44.824035	2.168316e-11
speechiness	8.828695e+03	1.0	30.150618	4.008466e-08
acousticness	6.008463e+03	1.0	20.519326	5.910862e-06
instrumentalness	4.337130e+03	1.0	14.811607	1.188848e-04
liveness	9.921465e+03	1.0	33.882507	5.874243e-09
valence	4.126982e+04	1.0	140.939375	1.755099e-32
tempo	1.367973e+03	1.0	4.671723	3.066580e-02
Residual	2.611073e+07	89170.0	NaN	NaN

The above F-test confirms that for the chosen model, all predictors included are indeed significant given all other predictors are in the model.

V. Genre Subset Regression - Kpop

Since our data was very noisy, it was difficult for our model to find any patterns. It also proved difficult to retrieve the influential points from the entire dataset. To improve the signal to noise ratio, and the efficiency of our model selection we narrowed our scope to a single genre. We chose Kpop as it had the highest average popularity score.

A, Feature selection

For genre, we were more selective about the features we chose. We decided to drop loudness because it is used in the calculation for energy, resulting in a high correlation (r=0.76). We dropped liveness because it is not an inherent characteristic of the song. Lastly, we dropped tracks with speechiness > 0.66, because these are considered to be spoken word tracks (i.e. podcasts). After

checking the VIF, we dropped the time signature again. In our initial model for predicting popularity for the kpop genre, we used duration_ms, explicit, danceability, energy, mode, speechiness, acousticness, instrumentalness, valence, and tempo.

	VIF score	features
0	593.756246	Intercept
1	1.127650	C(explicit)[T.True]
2	1.675111	C(key)[T.1]
3	1.700730	C(key)[T.2]
4	1.230569	C(key)[T.3]
5	1.583011	C(key)[T.4]
6	1.580039	C(key)[T.5]
7	1.504921	C(key)[T.6]
8	1.785014	C(key)[T.7]
9	1.446508	C(key)[T.8]
10	1.693601	C(key)[T.9]
11	1.489567	C(key)[T.10]
12	1.602050	C(key)[T.11]
13	1.094146	C(mode)[T.1]
14	5.932505	C(time_signature)[T.1]
15	43.570924	C(time_signature)[T.3]
16	55.991817	C(time_signature)[T.4]
17	10.337187	C(time_signature)[T.5]
18	1.109748	duration_ms
19	1.529390	danceability
20	2.649430	energy
21	1.173507	speechiness
22	2.478690	acousticness
23	1.165288	instrumentalness
24	1.624783	valence
25	1.093829	tempo

B, Model Diagnosis

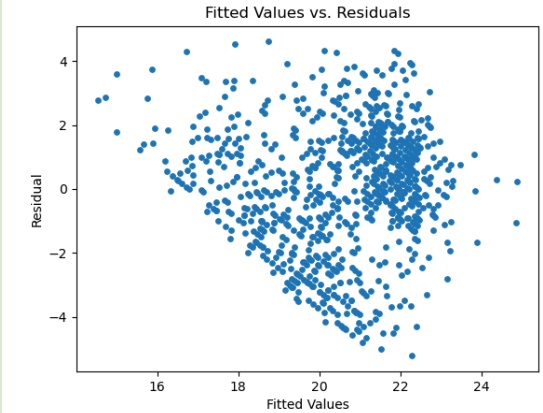
First, we dropped the influential points based on external studentized residual and Cook's distance. When investigating the presence of heteroscedasticity, the data passed the Bresuch-Pagan test and therefore we will not be using the robust standard error for t-test for predictor significance. Similarly, we were unable to alleviate the normality violations. However, the distribution was not significantly departed from normal and we have a large sample size, so we can continue.

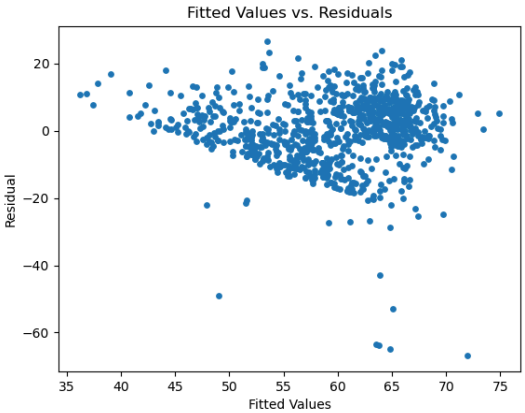
OLS Regression Results

Dep. Variable:	popularity	**R-squared:**	0.438
Model:	OLS	**Adj. R-squared:**	0.424
Method:	Least Squares	**F-statistic:**	31.48
Date:	Thu, 12 Oct 2023	**Prob (F-statistic):**	2.65e-91
Time:	16:56:19	**Log-Likelihood:**	-3046.3
No. Observations:	870	**AIC:**	6137.
Df Residuals:	848	**BIC:**	6242.
Df Model:	21		
Covariance Type:	nonrobust		

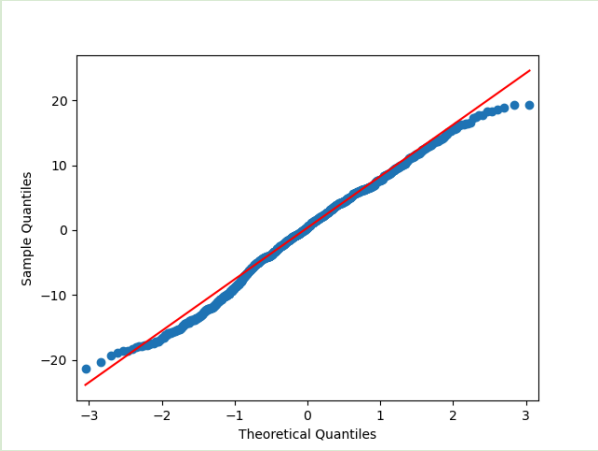
Omnibus:	12.765	**Durbin-Watson:**	1.304
Prob(Omnibus):	0.002	**Jarque-Bera (JB):**	10.572
Skew:	-0.193	**Prob(JB):**	0.00506
Kurtosis:	2.623	**Cond. No.**	5.94e+06

Without influential points	With influential points
BP test p-value: 0.01	BP test p-value: 0.01

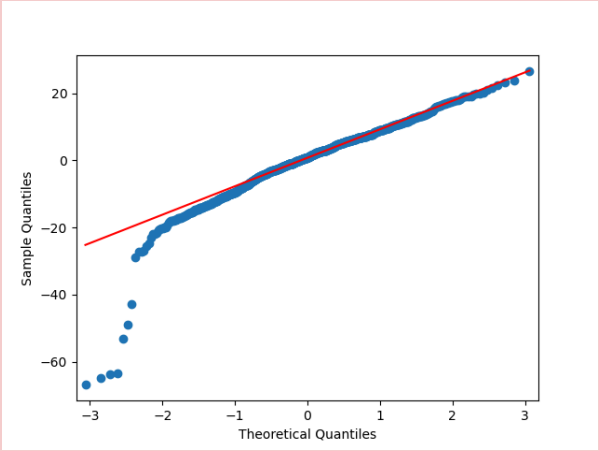




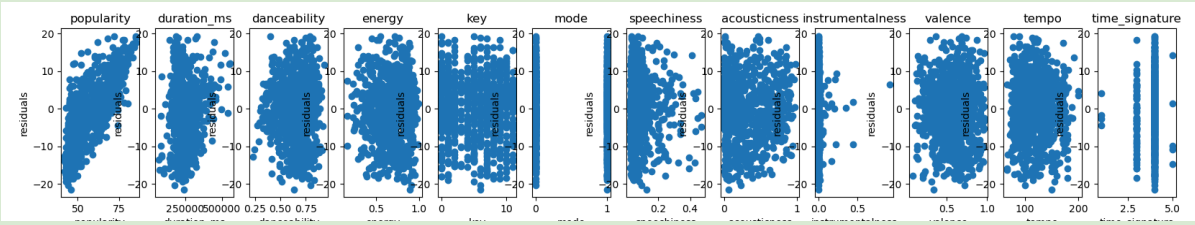
JB test p-value: 0.005



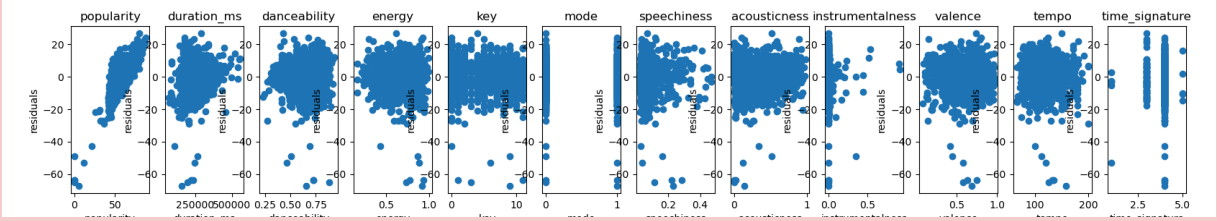
JB test p-value: 0.00



Without influential points:



With influential points



C, Model Selection

For the model selection process, we chose to use the best subset method using Mallow's Cp along with the adjusted R-squared. From this, we obtain a final model of popularity ~ duration_ms + explicit + danceability + mode + acousticness + instrumentalness + valence. We chose this because it had the best Mallow's Cp with 29.7983 (8 predicted parameters) and a relatively good adjusted R-squared of 0.42.

model	adj_r2	Cp
(duration_ms, explicit, danceability, mode, ac...	0.421882	29.798280
(duration_ms, explicit, danceability, energy, ...	0.422409	29.969143
(duration_ms, explicit, energy, mode, acoustic...	0.421368	30.584346
(duration_ms, explicit, mode, acousticness, in...	0.420500	30.939294
(duration_ms, explicit, danceability, mode, ac...	0.421307	31.651207

Models sorted by Cp ascending

D, Model Findings

From the t-tests we find that all of the p-values are below 0.05. This suggests that each predictor is significant in predicting popularity score given the other predictors are being used. Although the R^2 is relatively low, this can be attributed to the fact that our data set is composed of many subjective predictors, which are difficult to measure absolutely. The p-values, along with the R^2 value, could be interpreted as there is a possibility that there are even more significant predictors of popularity that we don't have access to.

	predictor	t-stat	p-val
0	int	39.799659	2.023265e-197
1	duration_ms	6.345336	3.581717e-10
2	explicit	-10.505207	2.254270e-24
3	danceability	-1.853893	6.409589e-02
4	mode	2.813626	5.010090e-03
5	acousticness	-13.785275	3.252674e-39
6	instrumentalness	-2.594451	9.635112e-03
7	valence	-2.608408	9.253926e-03

In terms of explicit model findings our coefficients for the model are as shown on the right. We use the top table p-values, as those are the robust standard error values.

	coef	std err	t	P>|t|
Intercept	80.2244	2.596	30.905	0.000
explicit[T.True]	9.0587	2.753	3.291	0.001
duration_ms	-5.437e-05	4.54e-06	-11.987	0.000
danceability	-3.9446	2.469	-1.598	0.110
energy	2.9241	2.188	1.336	0.182
mode	1.5459	0.557	2.775	0.006
acousticness	-14.5789	1.425	-10.229	0.000
instrumentalness	-20.2877	6.386	-3.177	0.002
valence	-4.7828	1.622	-2.949	0.003

Since all of the p-values are below 0.05, we can say that each of the predictors are significant in predicting popularity, given the other predictors are being used.

Interpretation of interesting coefficients:
On average, a song that is explicit is 9.0587 points more popular than a song that isn't.

Additionally, a song that is major will be 1.5459 points more popular than a song that is minor. For each millisecond longer a song is, it is on average 5.437e-05 points less popular.

VI. Conclusion

From the full model, it is clear that genre is a significant predictor of popularity. However, within each genre, listeners prefer different features. For example, for K-pop fans, the model shows that duration_ms, explicitness, danceability, mode, acousticness, instrumentalness, and valence significantly influence popularity.

Conversely, when considering all genres, all track features beside genre have a slight impact on the song's popularity. Therefore, the most effective model for predicting song popularity includes duration_ms, explicitness, danceability, energy, key, mode, acousticness, instrumentalness, liveness, valence, tempo, and track genre.

- Danceability: Songs that are easier to dance to attract a larger audience, potentially increasing their popularity.
- Explicitness: The presence of explicit content may appeal to a specific audience, impacting the song's popularity within that demographic.
- Energy: Lower energy tracks generally make them more popular in this model. This shows a trend for more ambient songs.
- Key: Different musical keys can evoke varied emotions; certain keys might resonate more with listeners, affecting the song's popularity.
- Mode: Major and minor modes convey different moods; listeners might prefer one over the other, influencing a song's popularity.
- Acousticness: Listeners favor modern music so more acoustic songs might appeal to listeners seeking a raw, natural sound, negatively impacting their popularity.
- Instrumentalness: Instrumental tracks might cater to listeners focusing on musical composition, affecting popularity within that niche audience.
- Liveness: Live recordings are less favored compared to studio records, influencing the song’s popularity negatively.
- Duration: Longer songs might have a different listener engagement compared to shorter ones, affecting popularity.
- Tempo: Faster or slower tempos can influence the emotional response; listeners might prefer songs with a specific tempo, affecting popularity.
- Track Genre: Different genres have distinct listener bases; songs belonging to top popular genres naturally gain higher popularity.

When considering all genres, these factors collectively contribute, albeit slightly, to a song's popularity. While the genre itself remains a significant predictor, these features add nuanced layers, catering to diverse listener preferences and impacting a song's overall appeal and popularity.

If we were to continue this study, it would be interesting to see which predictors would be significant for other genres. With this information, we could possibly isolate significant predictors specific to each genre and see how this relates to our full model.