ISYE 6414 Project:
Spotify music popularity analysis

Team #3 - Fall 2022

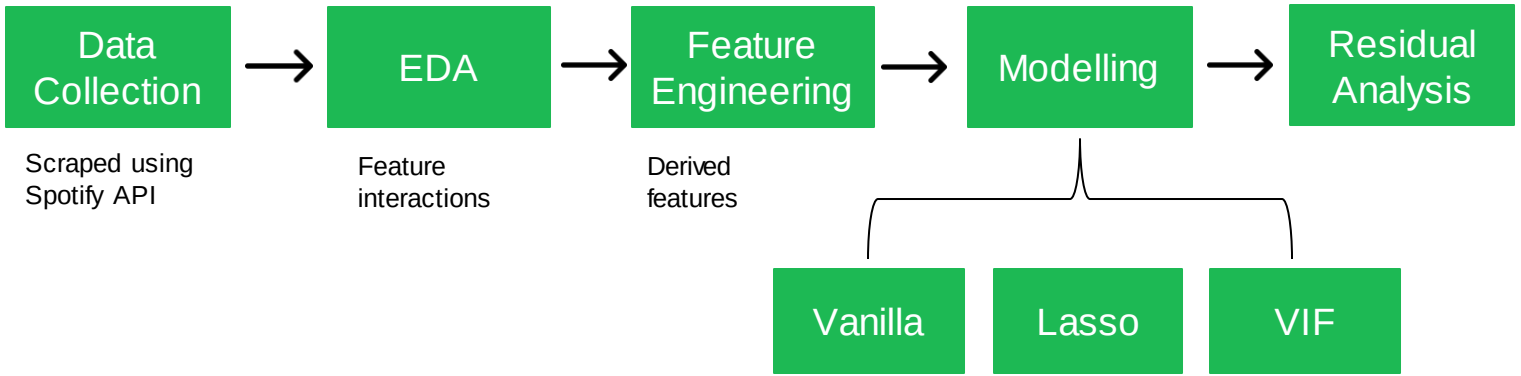
Kien Tran - ktran332@gatech.edu
Abhinav Arun - aarun60@gatech.edu
Anshit Verma - averma373@gatech.edu
Ashish Dhiman - adhiman9@gatech.edu
Saksham Arora - sarora320@gatech.edu

Problem statement

There's hardly any individual on the planet that does not enjoy music. Being such an integral part of human civilization, even conservative estimates value the **Music Industry at upwards of $25B**. Hence as part of our project for this course, we want to dive deeper into the dynamics of music popularity, and the various factors driving it. We want to explore complex **features associated with music** and what **external factors contribute towards the traction of a song**. In other words, we want to build a model to predict the traction of a song and find out different factors that affect such traction.

Our goal is to build a system that would not only allow us to **predict the popularity of a song** but at the same time help us **identify if there are certain factors that increase the propensity of a song** to be popular. Our system would thus have **both a predictive as well as descriptive** use case. While our system should recognize expected patterns, like a few key artists (say Drake, etc.), we also hope to uncover some lesser-known patterns, say whether songs composed in a few particular scales tend to be more popular.

Methodology: Pipeline



Data Collection (1/2)

Top 10 Tracks for Top Artists

Using Client Credential Authorization Flow

Spotify Database

Number of tracks = ~5000
Number of Artists = 729

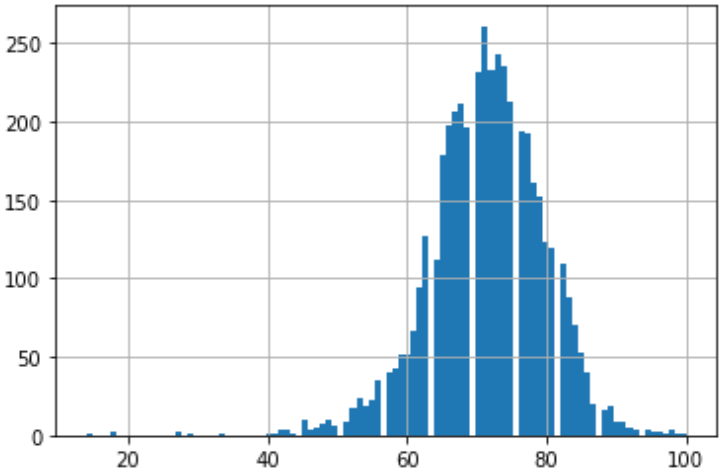
	name	album	artist	release_date	length	popularity	danceability	acousticness	energy
323	Perfect	Made In The A.M. (Deluxe Edition)	One Direction	2015-11-13	230333	79	0.647	0.0598	0.823
385	Without Me	Manic	Halsey	2020-01-17	201660	80	0.752	0.2970	0.488
451	Happier	Happier	Marshmello	2018-08-17	214289	84	0.687	0.1910	0.792
466	Numb	Meteora	Linkin Park	2003-03-24	185586	84	0.496	0.0046	0.863
614	West Coast	West Coast	OneRepublic	2022-02-25	192947	77	0.685	0.3170	0.699
...	...	...	...	...	...	...	...	...	...
4497	Forever	Beautiful Mind	Rod Wave	2022-08-12	214185	69	0.626	0.5660	0.524
4527	HOT	SEVENTEEN 4th Album 'Face the Sun'	SEVENTEEN	2022-05-27	197586	81	0.765	0.0539	0.777
4537	September	The Best Of Earth, Wind & Fire Vol. 1	Earth, Wind & Fire	1978-11-23	215093	84	0.697	0.1680	0.832
4544	Fantasy	All 'N All	Earth, Wind & Fire	1977-11-21	277413	59	0.608	0.3230	0.745
4572	Invincible	Spider-Man: Into the Spider-Verse (Soundtrack ...	Various Artists	2018-12-14	196386	61	0.724	0.1410	0.600

Data Collection (2/2)

Variable Name	Type	Description
Name	Qualitative	Name of Track
Album	Qualitative	Name of Album
Artist	Qualitative	Name of Artist
Length	Qualitative	Duration of Track
Danceability	Quantitative	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 the least danceable and 1.0 is the most danceable.
Acousticness	Quantitative	A confidence measure from 0.0 to 1.0 of whether the track is acoustic. 1.0 represents high confidence the track is acoustic.
Energy	Quantitative	Energy is a measure from 0.0 to 1.0 and represents a perceptual measure of intensity and activity. Typically, energetic tracks feel fast, loud, and noisy. For example, death metal has high energy, while a Bach prelude scores low on the scale. Perceptual features contributing to this attribute include dynamic range, perceived loudness, timbre, onset rate, and general entropy.
Instrumentalness	Quantitative	Predicts whether a track contains no vocals. "Ooh" and "aah" sounds are treated as instrumental in this context. Rap or spoken word tracks are clearly "vocal". The closer the instrumentalness value is to 1.0, the greater likelihood the track contains no vocal content.
Liveness	Quantitative	Detects the presence of an audience in the recording. Higher liveness values represent an increased probability that the track was performed live. A value above 0.8 provides strong likelihood that the track is live.
Loudness	Quantitative	The overall loudness of a track in decibels (dB). Loudness values are averaged across the entire track and are useful for comparing relative loudness of tracks. Loudness is the quality of a sound that is the primary psychological correlate of physical strength (amplitude). Values typically range between -60 and 0 db.
Speechiness	Quantitative	Speechiness detects the presence of spoken words in a track. The more exclusively speech-like the recording (e.g. talk show, audio book, poetry), the closer to 1.0 the attribute value. Values above 0.66 describe tracks that are probably made entirely of spoken words. Values between 0.33 and 0.66 describe tracks that may contain both music and speech, either in sections or layered, including such cases as rap music. Values below 0.33 most likely represent music and other non-speech-like tracks.
Tempo	Quantitative	The overall estimated tempo of a track in beats per minute (BPM). In musical terminology, tempo is the speed or pace of a given piece and derives directly from the average beat duration.
Time signature	Qualitative	An estimated time signature. The time signature (meter) is a notational convention to specify how many beats are in each bar (or measure). The time signature ranges from 3 to 7 indicating time signatures of "3/4", to "7/4". >= 3<= 7

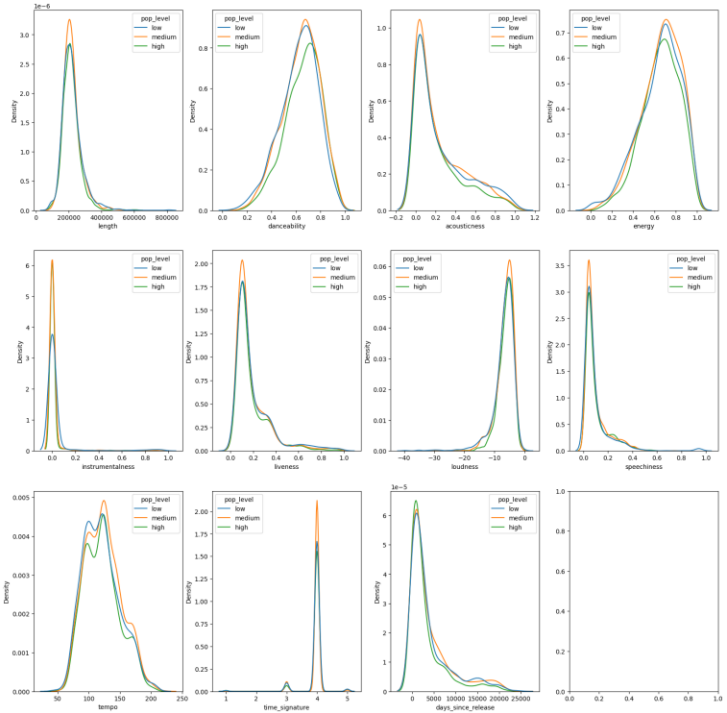
Exploratory Data Analysis (EDA) (1/3)

Distribution of Popularity



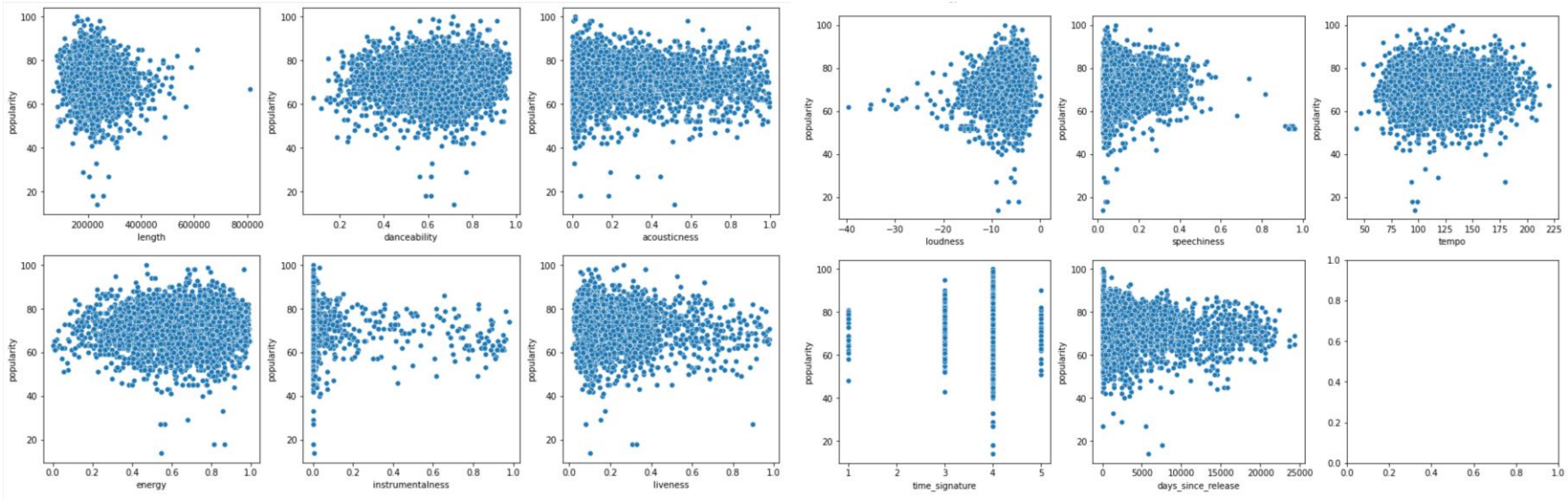
Popularity (Target Variable) is Normally Distributed with a mean of around 71

Distribution of Features



Exploratory Data Analysis (EDA) (2/3)

We plotted Popularity (target) against 11 song attributes => There wasn't good correlations between popularity and the song attributes.



Exploratory Data Analysis (EDA) (3/3)

The lack of predicting power was further confirmed by the quick regression model reaching an R-squared of only 0.03
However, adding another attribute derived from popularity of the same artist helps increase R-squared significantly (to 0.44)
=> Feature engineering was going to be pivotal, we did more EDA to form hypotheses that support feature engineering

```
Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)       7.691e+01  2.118e+00  36.309  < 2e-16 ***
length           -1.686e-06  2.223e-06  -0.758   0.4482
danceability      3.736e+00  8.798e-01   4.247 2.21e-05 ***
acousticness     -1.173e+00  6.036e-01  -1.943   0.0521 .
energy           -5.812e+00  1.065e+00  -5.457 5.11e-08 ***
instrumentalness -1.866e+00  1.182e+00  -1.578   0.1145
liveness         -2.264e+00  7.926e-01  -2.857   0.0043 **
loudness          4.460e-01  6.186e-02   7.210 6.50e-13 ***
speechiness      -1.822e+00  1.207e+00  -1.510   0.1312
tempo             9.132e-03  4.269e-03   2.139   0.0325 *
time_signature   -2.541e-01  4.031e-01  -0.630   0.5285
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 8.081 on 4565 degrees of freedom
Multiple R-squared:  0.03069,   Adjusted R-squared:  0.02856
F-statistic: 14.45 on 10 and 4565 DF,  p-value: < 2.2e-16
```

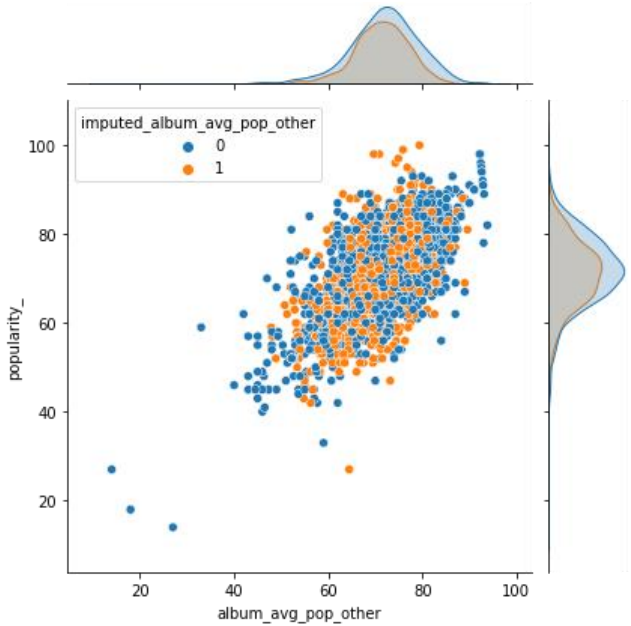
```
Coefficients:
                            Estimate Std. Error t value Pr(>|t|)
(Intercept)                1.173e+01  1.925e+00   6.092  1.2e-09 ***
length                    -2.700e-06  1.708e-06  -1.581  0.11401
danceability               1.853e+00  6.611e-01   2.804  0.00507 **
acousticness               1.435e-01  4.528e-01   0.317  0.75136
energy                    -1.271e+00  8.146e-01  -1.561  0.11869
instrumentalness          -3.961e-01  9.146e-01  -0.433  0.66498
liveness                  -9.804e-01  5.956e-01  -1.646  0.09982 .
loudness                   1.604e-01  5.000e-02   3.208  0.00135 **
speechiness               -1.728e+00  9.179e-01  -1.882  0.05984 .
tempo                      2.913e-03  3.152e-03   0.924  0.35541
time_signature            -4.668e-01  2.912e-01  -1.603  0.10897
mean_other_song_popularity 8.759e-01  1.476e-02  59.333  < 2e-16 ***
days_since_release         7.271e-05  2.244e-05   3.240  0.00121 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

 Residual standard error: 6.134 on 4806 degrees of freedom
   (181 observations deleted due to missingness)
 Multiple R-squared:  0.4425,    Adjusted R-squared:  0.4411
 F-statistic: 317.8 on 12 and 4806 DF,  p-value: < 2.2e-16
```

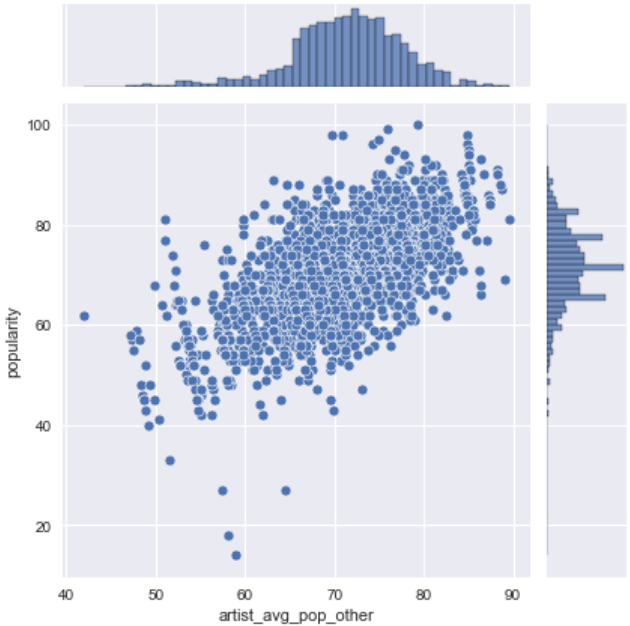
Feature Engineering (1/2)

Because the original features did not provide enough predictive power, we have accentuated by our data with feature engineering

Average Popularity of Album



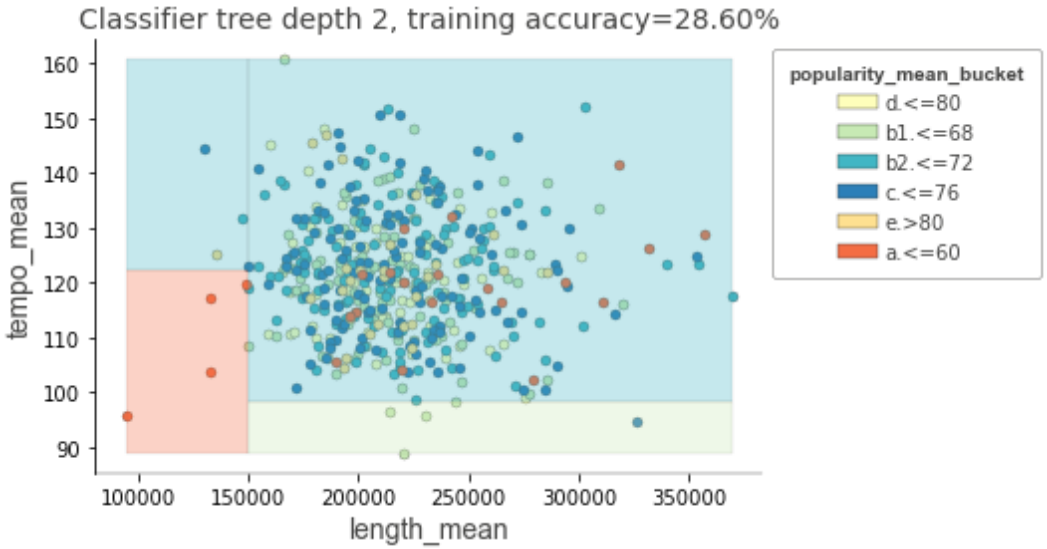
Average Popularity of Artist



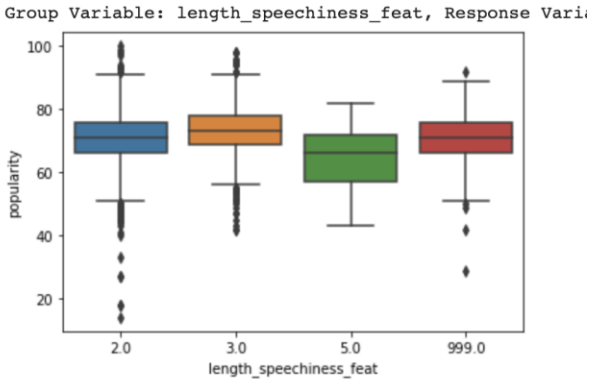
Feature Engineering (2/2)

We have created categorical features basis the spatial analysis of artists along features like instrumentalness, liveness etc.

Spatial Analysis



Tukey Test



Multiple Comparison of Means - Tukey HSD, FWER=0.01

group1	group2	meandiff	p-adj	lower	upper	reject
2.0	3.0	2.342	0.001	1.4647	3.2193	True
2.0	5.0	-7.2288	0.001	-10.0771	-4.3806	True
2.0	999.0	-0.2646	0.9	-1.7836	1.2544	False
3.0	5.0	-9.5709	0.001	-12.4813	-6.6604	True
3.0	999.0	-2.6066	0.001	-4.2391	-0.974	True
5.0	999.0	6.9643	0.001	3.8007	10.1279	True

Modelling: Vanilla Model

After feature engineering, our baseline Regression Model shows up a R-squared of 68.5%
However, a lot of our predictors have high p-values and therefore are not significant for regression.

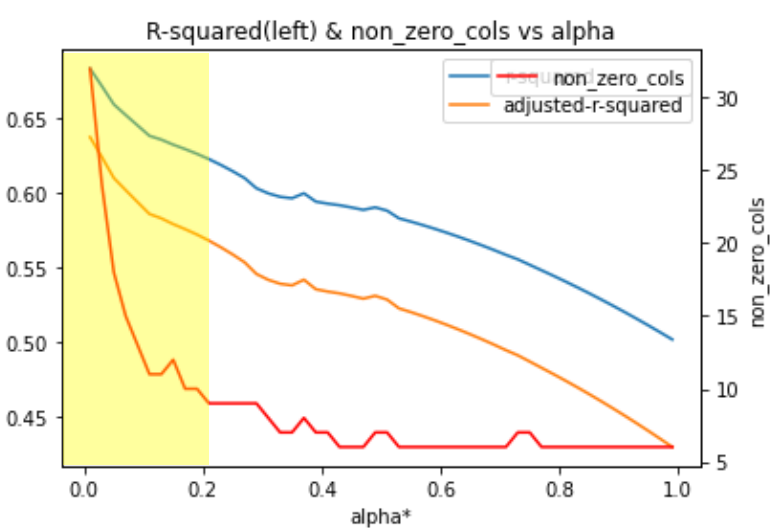
	coef	std err	t	P>|t|	[0.025	0.975]
const	20.1692	5.547	3.636	0.000	9.236	31.102
length	-6.4982	4.848	-1.341	0.181	-16.053	3.057
danceability	0.0792	2.395	0.033	0.974	-4.642	4.800
acousticness	-0.7187	2.850	-0.252	0.801	-6.336	4.898
energy	-3.7760	3.047	-1.239	0.217	-9.782	2.230
instrumentalness	12.6152	7.998	1.577	0.116	-3.149	28.379
liveness	1.5767	2.517	0.627	0.532	-3.384	6.537
loudness	11.7989	6.147	1.920	0.056	-0.317	23.915
speechiness	2.3517	4.026	0.584	0.560	-5.584	10.287
tempo	-7.1755	2.367	-3.032	0.003	-11.841	-2.510
days_since	2.8295	1.837	1.541	0.125	-0.790	6.450
popularity_meta	17.1241	4.136	4.140	0.000	8.972	25.276
followers	7.4146	12.404	0.598	0.551	-17.034	31.863
popularity_count	3.7524	5.612	0.669	0.504	-7.309	14.814
popularity_count_album	-1.5721	3.144	-0.500	0.618	-7.770	4.626
artist_avg_pop_other	-6.0799	7.211	-0.843	0.400	-20.294	8.134
album_avg_pop_other	30.5589	6.758	4.522	0.000	17.238	43.880
imputed_album_avg_pop_other	1.6590	1.588	1.045	0.297	-1.470	4.788

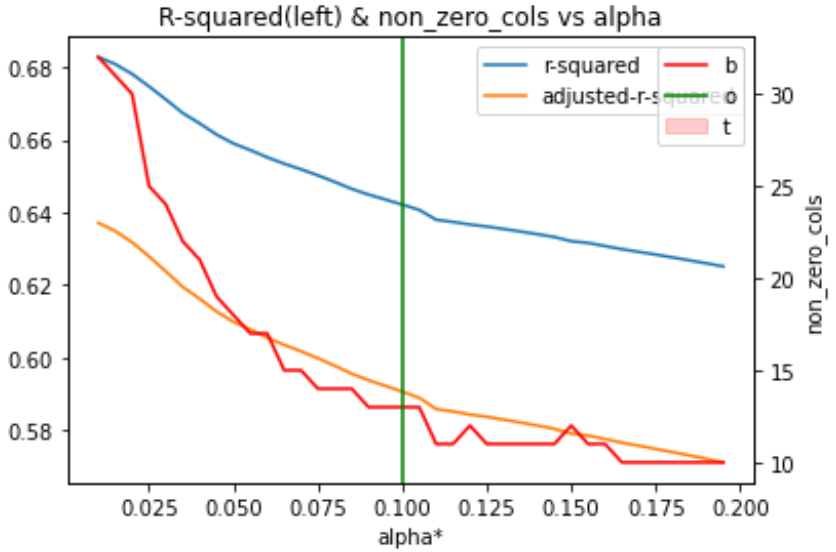
OLS Regression Results

Dep. Variable:	popularity_	**R-squared:**	0.685
Model:	OLS	**Adj. R-squared:**	0.640
Method:	Least Squares	**F-statistic:**	15.11
Date:	Sun, 27 Nov 2022	**Prob (F-statistic):**	1.07e-38
Time:	18:37:09	**Log-Likelihood:**	-768.27
No. Observations:	247	**AIC:**	1601.
Df Residuals:	215	**BIC:**	1713.
Df Model:	31		
Covariance Type:	nonrobust		

Modelling: LASSO Regression

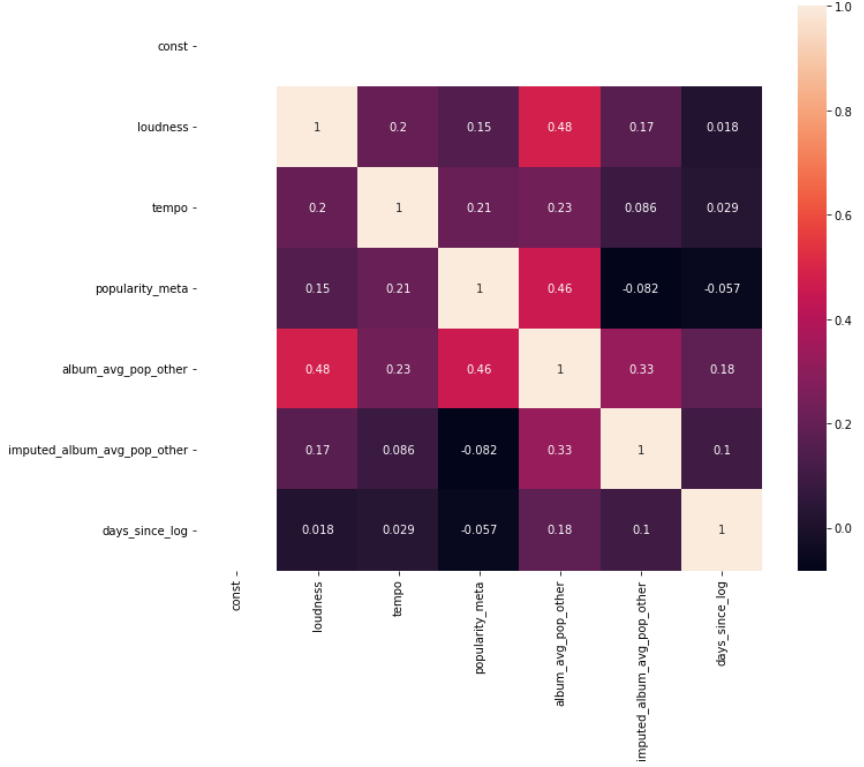
We have tried LASSO Regression for feature selection. After trying a range of values, we have chosen alpha of 0.1.
After employing LASSO we have dropped number of predictors in our model from 39 to 13.





Modelling: VIF

To handle multicollinearity, we have looked at Variance Inflation Factors, and added interaction terms to handle it.

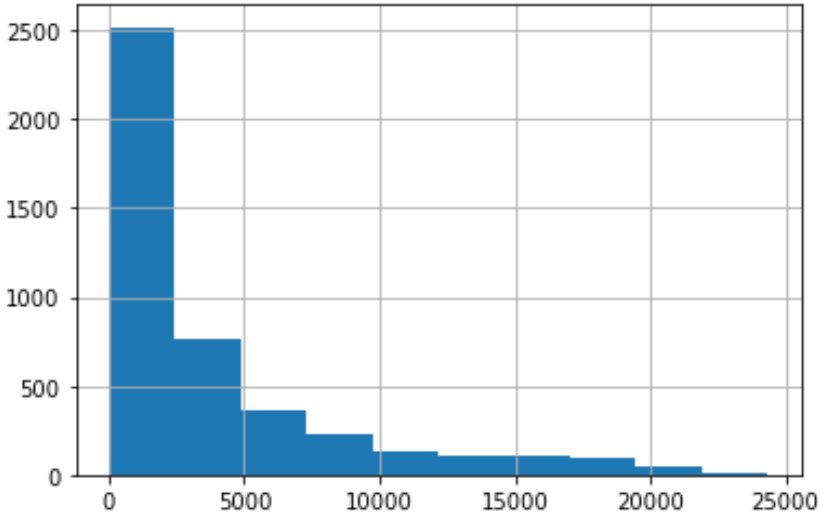


	VIF	variable
0	272.969828	const
4	2.006594	album_avg_pop_other
3	1.448989	popularity_meta
1	1.349130	loudness
5	1.213831	imputed_album_avg_pop_other
2	1.089849	tempo
6	1.071862	days_since_log

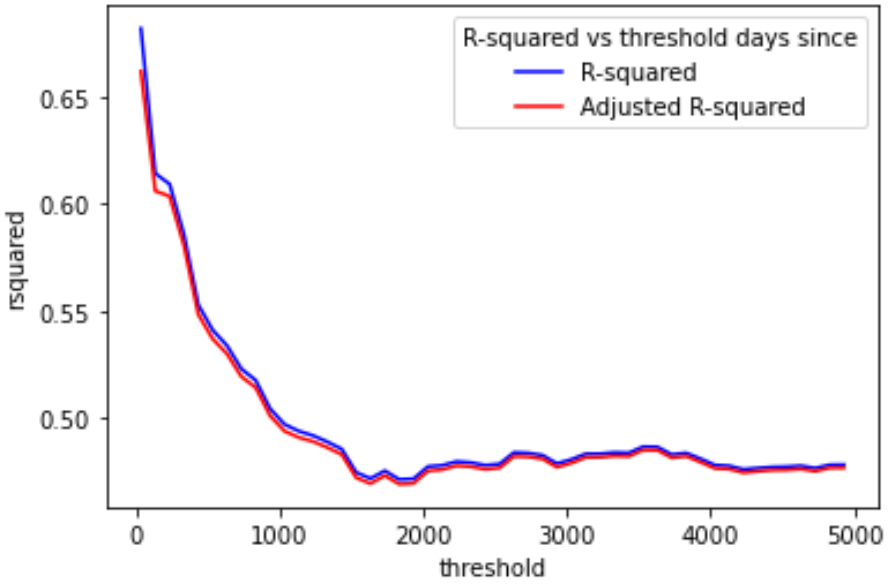
Temporal Effects

Popularity has temporal factors involved and we see the explanation power of features is higher for recently released songs.

Distribution of Days Since Release



R-squared vs Recent data



Final Model

The final model achieved after the feature engineering and feature selection can be seen. Every coefficient is significant and helps explain the variability of the dependent variable.

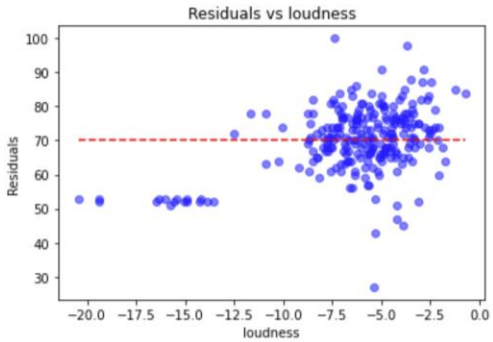
OLS Regression Results

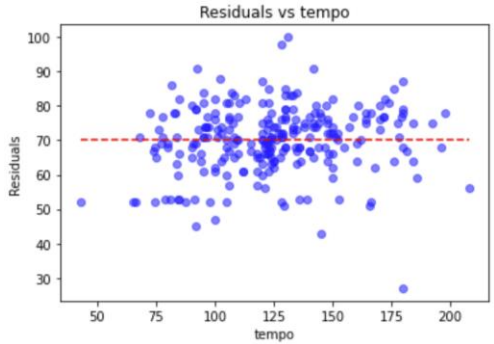
Dep. Variable:	popularity_	**R-squared:**	0.662
Model:	OLS	**Adj. R-squared:**	0.649
Method:	Least Squares	**F-statistic:**	51.58
Date:	Sun, 27 Nov 2022	**Prob (F-statistic):**	5.98e-51
Time:	23:29:42	**Log-Likelihood:**	-777.11
No. Observations:	247	**AIC:**	1574.
Df Residuals:	237	**BIC:**	1609.
Df Model:	9		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	20.8900	6.829	3.059	0.002	7.437	34.343
loudness	0.3515	0.173	2.035	0.043	0.011	0.692
tempo	-0.0350	0.013	-2.717	0.007	-0.060	-0.010
imputed_album_avg_pop_other	2.8240	0.776	3.637	0.000	1.295	4.354
danceability_speechiness_feat_3.0	-1.1408	0.820	-1.391	0.166	-2.756	0.475
length_speechiness_feat_999.0	4.1716	2.095	1.991	0.048	0.044	8.299
instrumentalness_time_signature_feat_2.0	8.4704	5.801	1.460	0.146	-2.958	19.899
instrumentalness_time_signature_feat_999.0	4.1716	2.095	1.991	0.048	0.044	8.299
length_tempo_feat_2.0	3.4963	2.340	1.494	0.137	-1.115	8.107
length_tempo_feat_999.0	4.1716	2.095	1.991	0.048	0.044	8.299
days_since_log	0.7040	0.163	4.324	0.000	0.383	1.025
popularity_meta_album_avg_pop_other	0.0076	0.001	14.838	0.000	0.007	0.009

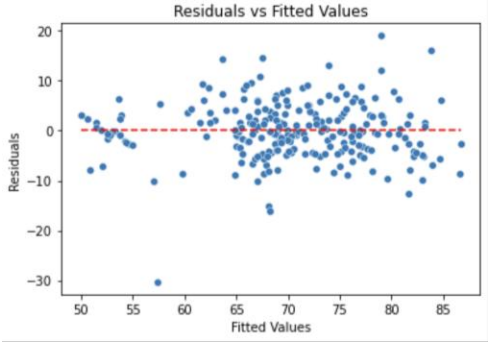
Residual Analysis

Linearity Assumption

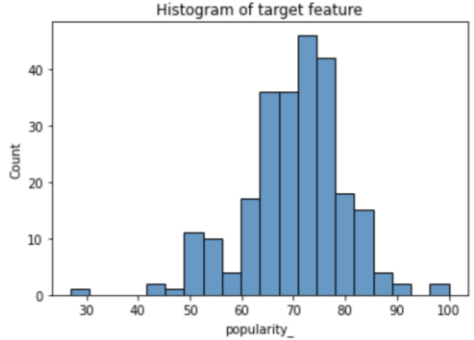


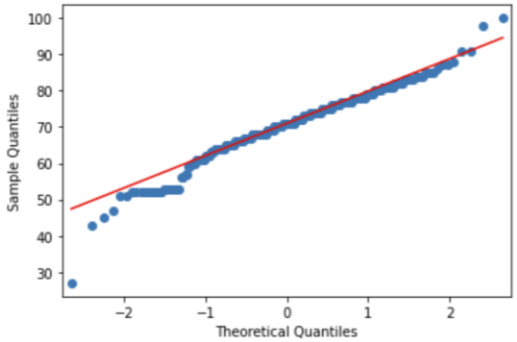


Constant Variance Assumption



Normality Assumption





Interpretation

- Intuitively, Artist and Album Popularity has a lot to do with song popularity
- Loudness has a positive relation to popularity
- Higher popularity is related with higher Number of Days Since Release
- Tempo has a slight negative relation with song popularity

	coef	std err	t	P>|t|	[0.025	0.975]
const	20.8900	6.829	3.059	0.002	7.437	34.343
loudness	0.3515	0.173	2.035	0.043	0.011	0.692
tempo	-0.0350	0.013	-2.717	0.007	-0.060	-0.010
imputed_album_avg_pop_other	2.8240	0.776	3.637	0.000	1.295	4.354
danceability_speechiness_feat_3.0	-1.1408	0.820	-1.391	0.166	-2.756	0.475
length_speechiness_feat_999.0	4.1716	2.095	1.991	0.048	0.044	8.299
instrumentalness_time_signature_feat_2.0	8.4704	5.801	1.460	0.146	-2.958	19.899
instrumentalness_time_signature_feat_999.0	4.1716	2.095	1.991	0.048	0.044	8.299
length_tempo_feat_2.0	3.4963	2.340	1.494	0.137	-1.115	8.107
length_tempo_feat_999.0	4.1716	2.095	1.991	0.048	0.044	8.299
days_since_log	0.7040	0.163	4.324	0.000	0.383	1.025
popularity_meta_album_avg_pop_other	0.0076	0.001	14.838	0.000	0.007	0.009

Conclusion and potential future works

From our analysis, the following insights are clear

- The popularity of a song are heavily influenced by the artist, and to be more specific, the popularity of other songs in the same album
- The influence is stronger for the recently-released songs (R-squared up to 0.65). After 3 years since release day, the influence seems to plateau.
- Among musical attributes, `loudness`, and `tempo` are the two most important factors in explaining a song's popularity, given that non-musical attributes like album popularity and release date remain fixed.

Potential future works

- We can improve our modeling by introducing a logistic regression model on top of our regression model which can predict the probability whether a song will get popular or not (decided by a threshold). This will help us improve the data we use for our linear regression model and potentially increase the model performance.
- For this project, we were only able to collect a snapshot of the data. More analysis on the time-series nature of popularity and bucketing based on genre might be a potential improvement for future projects (when multiple snapshots of the data are collected).

Appendix

- https://www.statsmodels.org/dev/generated/statsmodels.regression.linear_model.OLS.fit_regularized.html
- https://www.statisticshowto.com/lasso-regression/#:~:text=What%20is%20Lasso%20Regression%3F,i.e.%20models%20with%20fewer%20parameters).
- https://towardsdatascience.com/feature-engineering-for-machine-learning-3a5e293a5114
- https://www.statisticshowto.com/residual-plot/