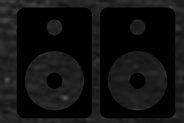






Analyzing Song Popularity Using Spotify's Audio Features

Jennie Sun



Spotify Dataset 1921-2020, 160k+ Tracks

Continuous Variables

- *acousticness*
- *danceability*
- *duration_ms*
- *energy*
- *instrumentalness*
- *liveness*
- *loudness*
- *speechiness*
- *tempo*
- *valence*
- year

Categorical Variables

- *explicit*
- *key*
- *mode*

Other Variables

- *artists*
- *id*
- *name*
- *release_date*

Response Variable

popularity (1 – 100)

⇩

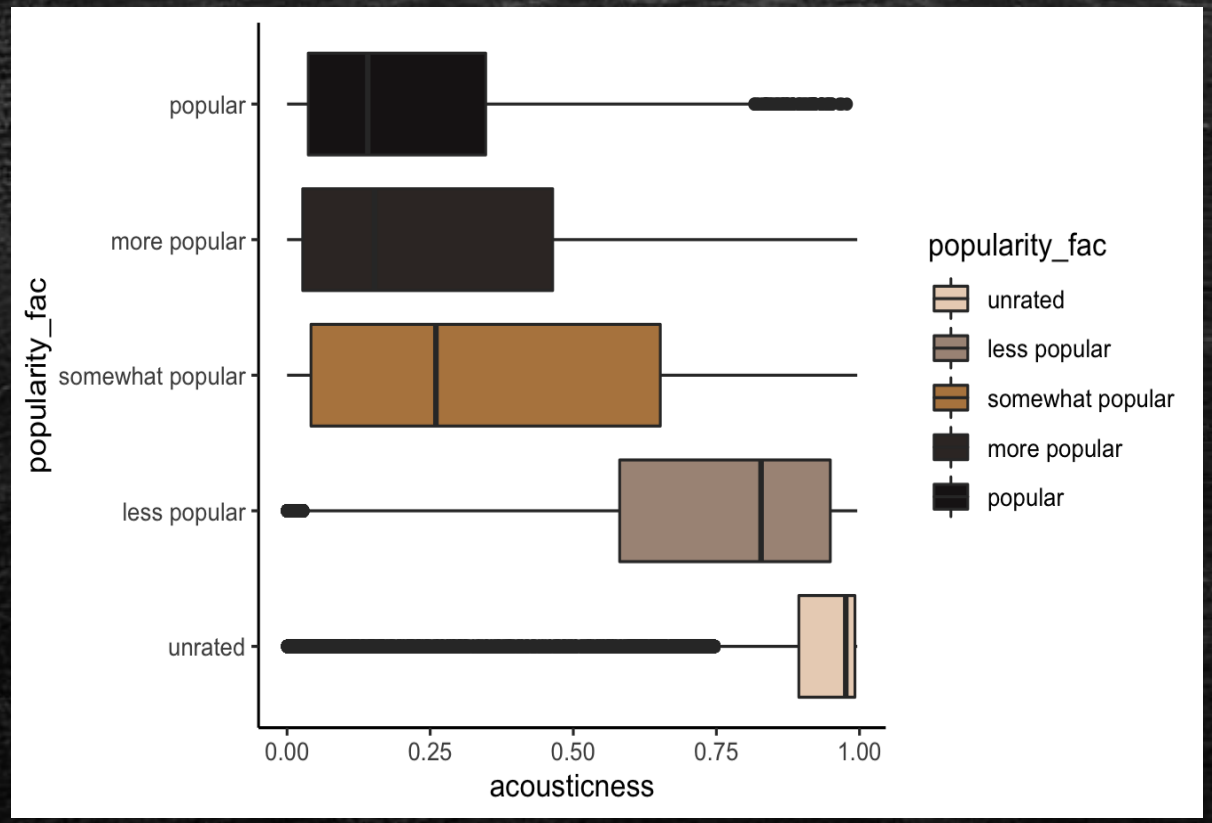
popularity_fac (5 levels)

- Unrated (0)
- Less popular (1-24)
- Somewhat popular (25-49)
- More popular (50-74)
- Popular (75-100)

Research Questions

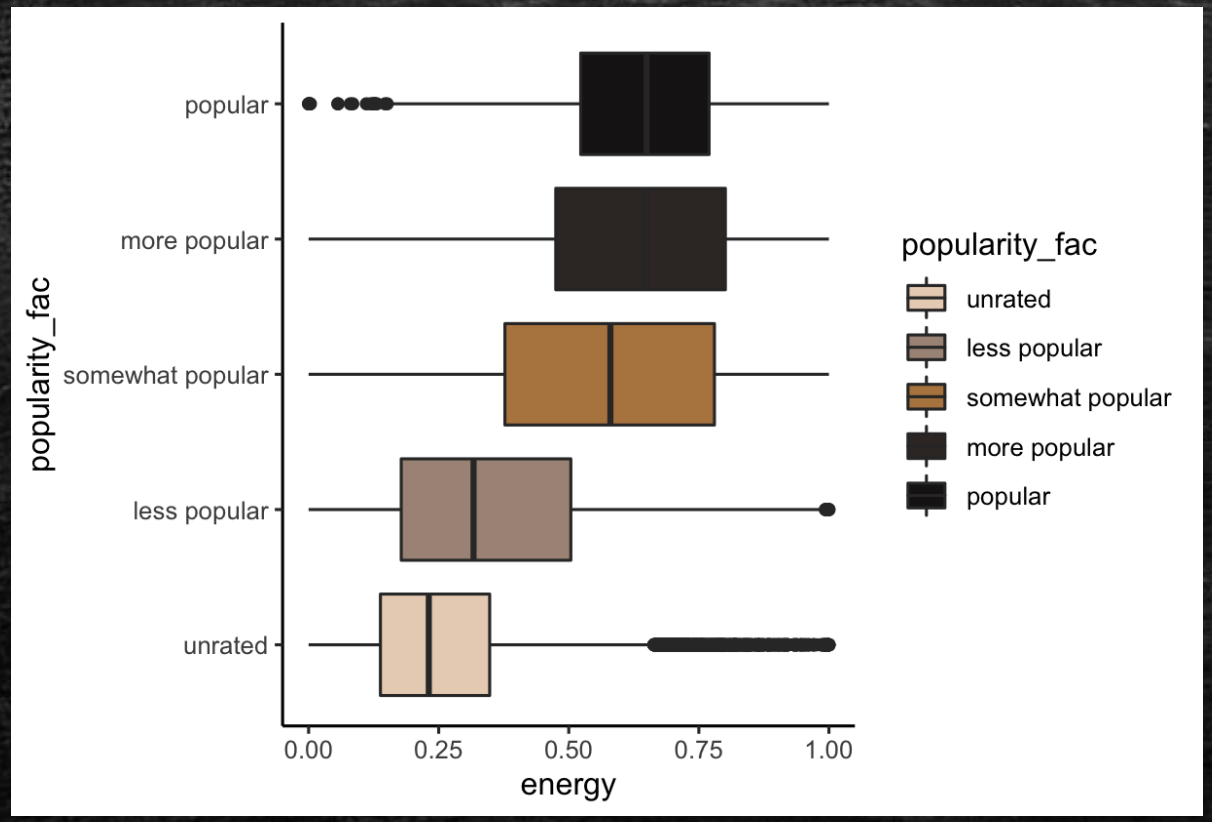
- What are the potential determinants for song popularity on Spotify?
- Are certain audio features affect popularity more than others? (if so, to what extent?)
- Are these features different for songs in different popularity levels?
- Any interesting interactions between audio features that affect song popularity?

Exploratory Data Analysis - Acousticness



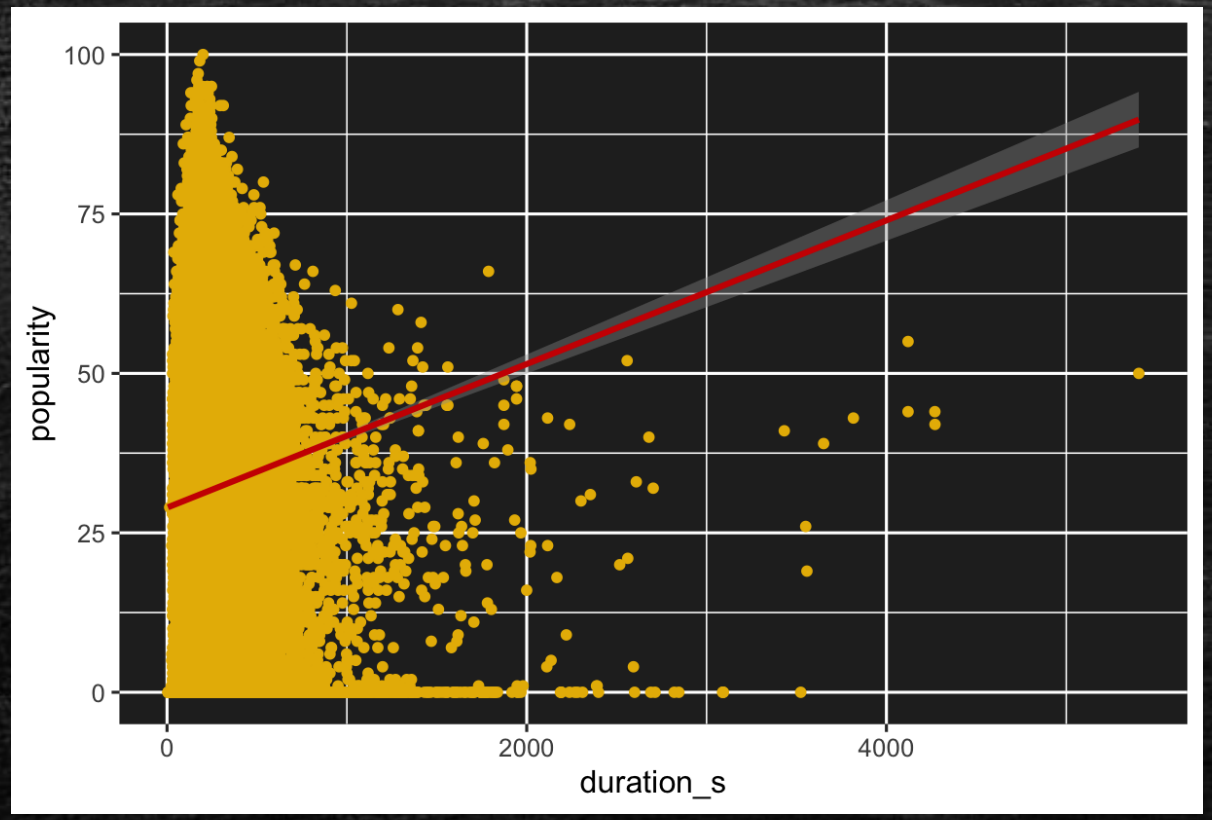
Inverse relationship between acousticness and popularity levels

Exploratory Data Analysis - Energy



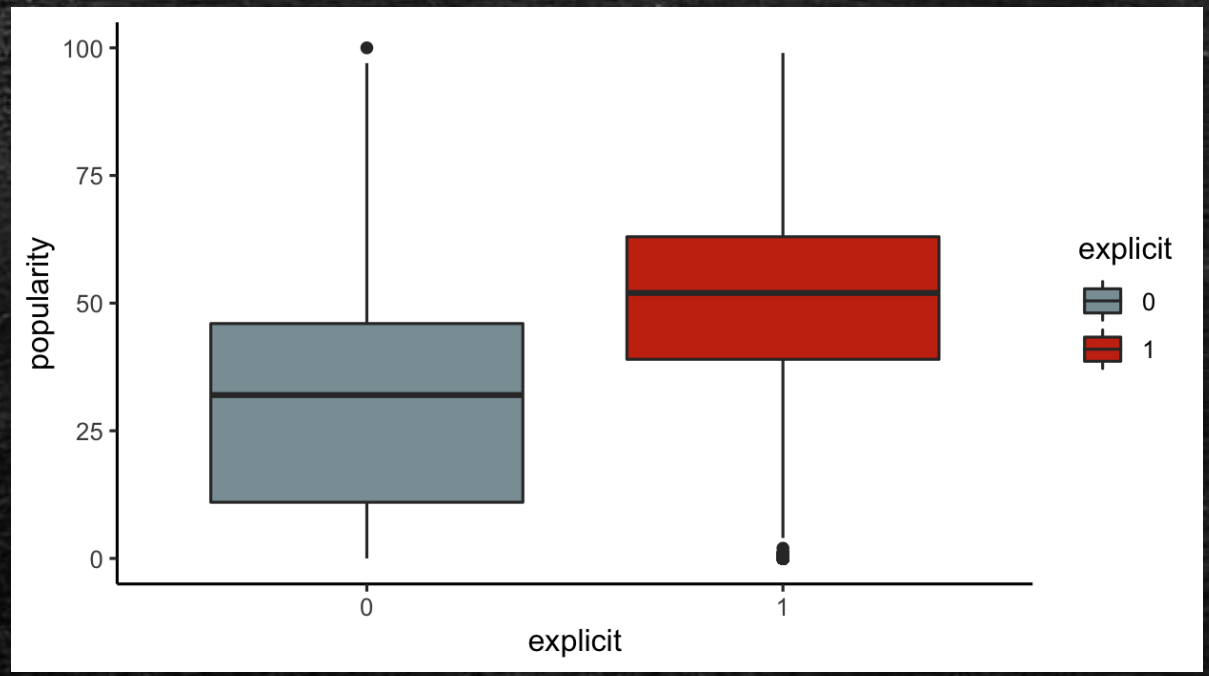
Positive relationship between energy and popularity levels

Exploratory Data Analysis - Duration



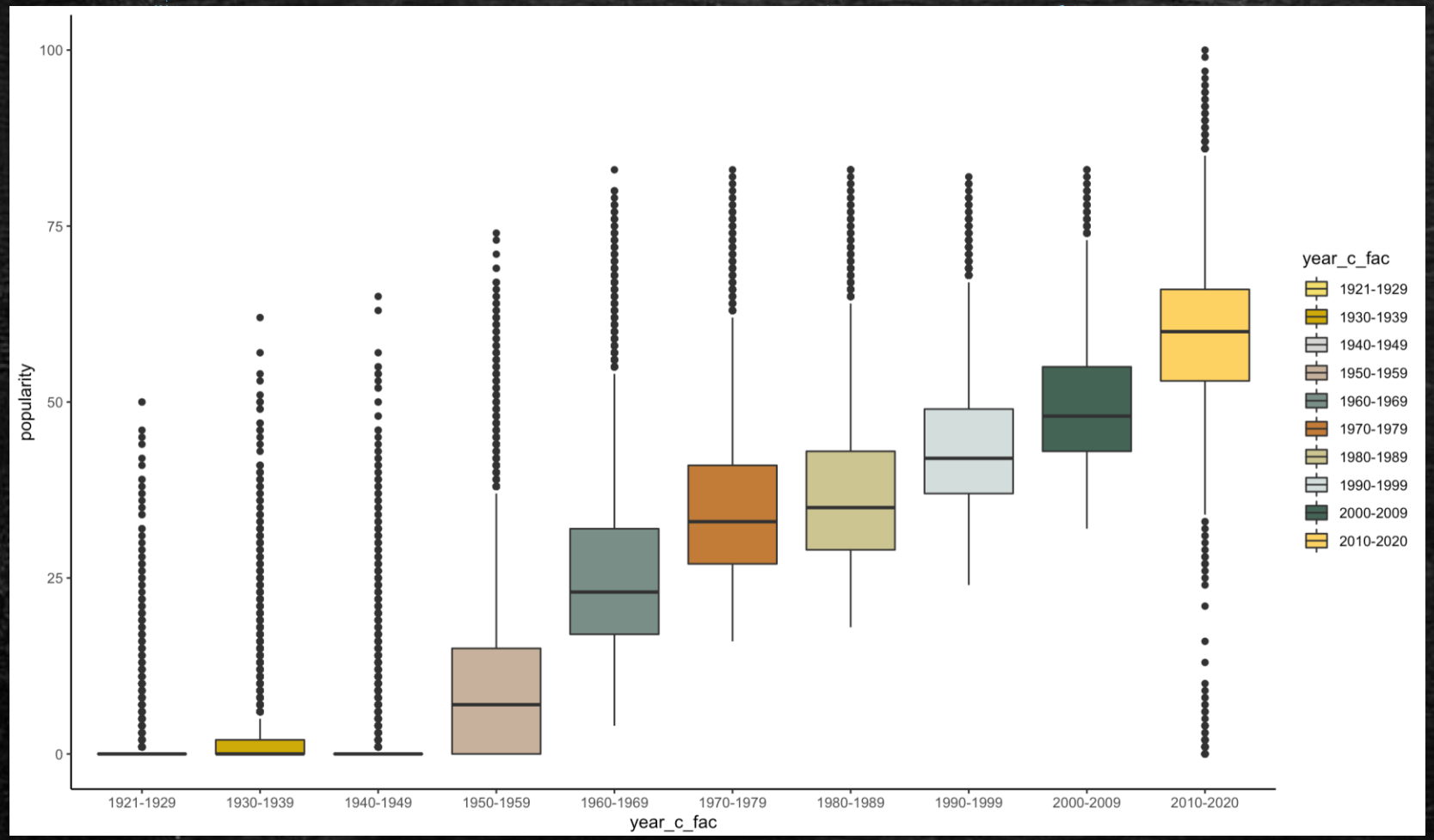
Positive relationship between duration and popularity scores

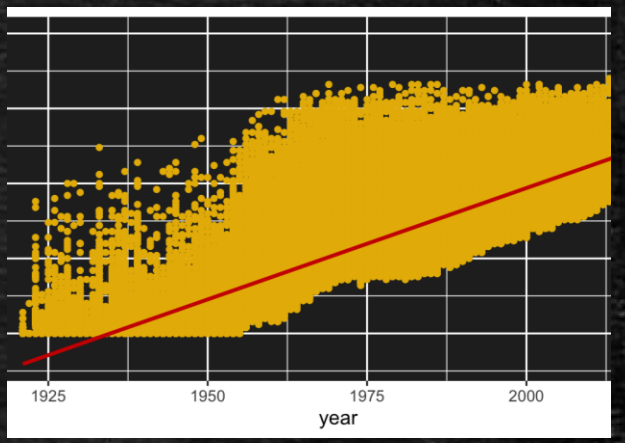
Exploratory Data Analysis - Explicit



A song seems more likely to receive a higher popularity score if it is marked as explicit

Exploratory Data Analysis - Year





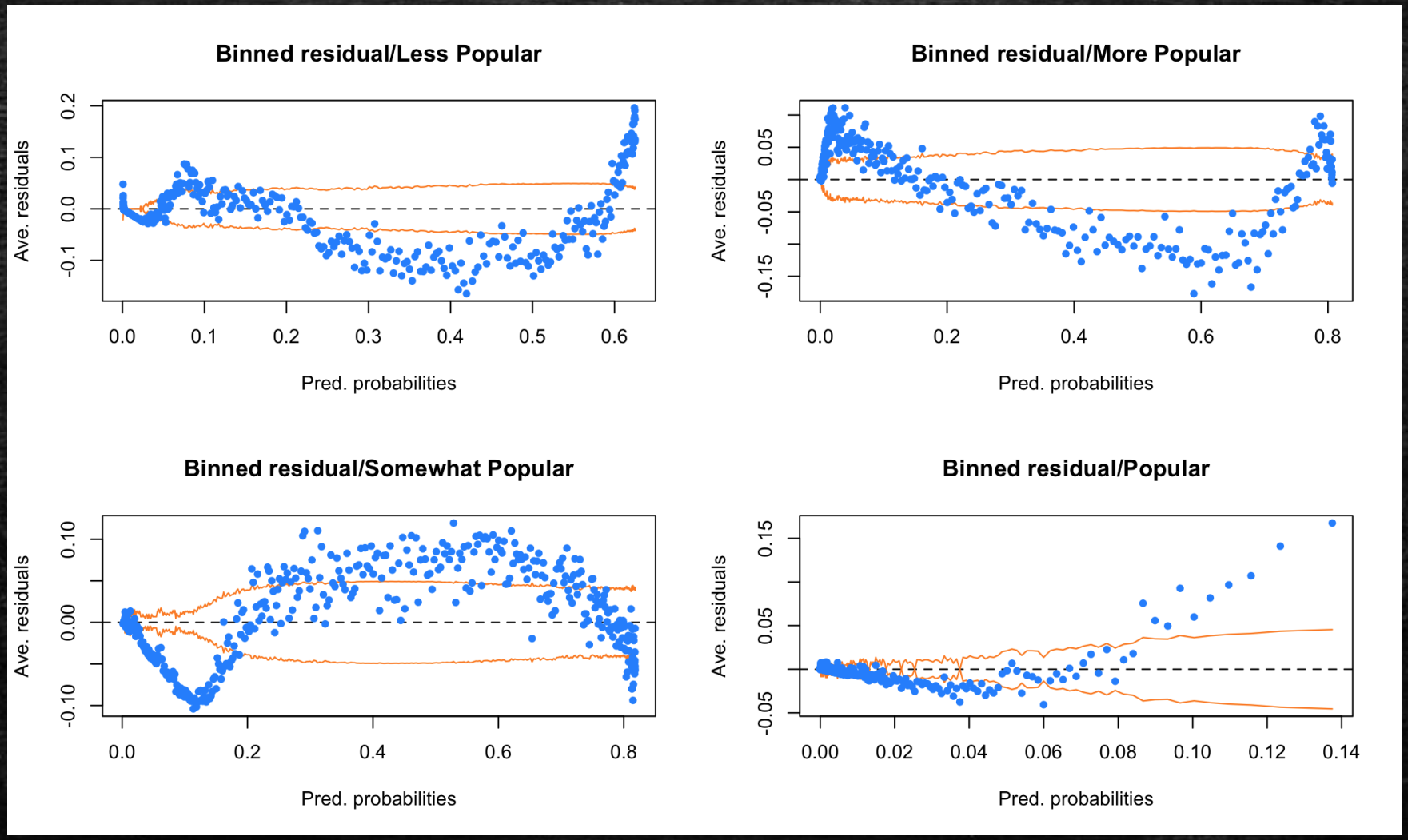
Newer songs seem to be more popular than older songs

Proportional Odds Model 1

```
model_1 <- polr(popularity_fac ~ acousticness + danceability
  + energy + instrumentalness + speechiness + duration_s +
  loudness + explicit + year_c)
```

	Residual Deviance	**AIC**
model_1	253660	253686

Proportional Odds Model 1 Diagnostics

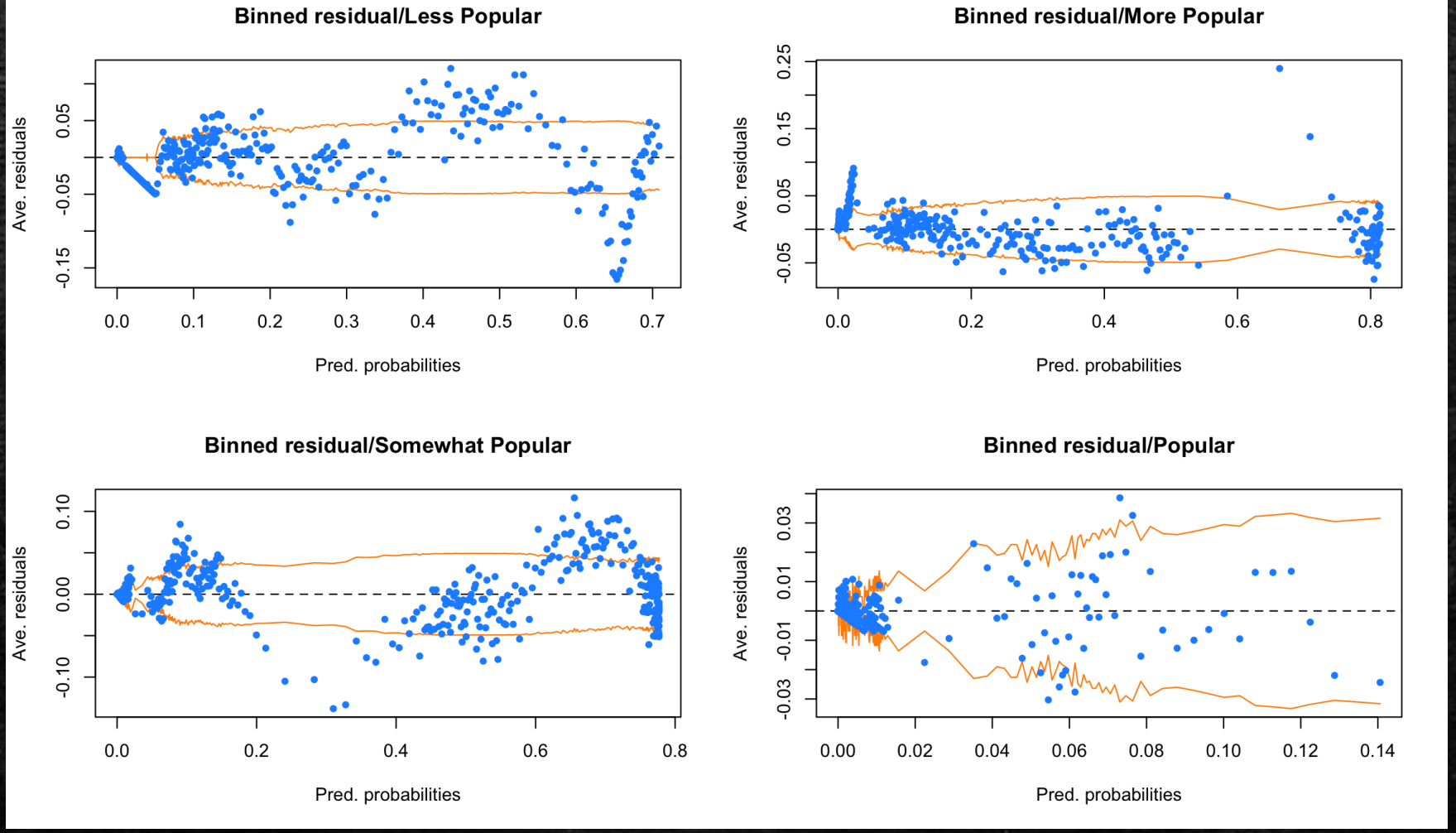


Proportional Odds Model 2

```
model_2 <- polr(popularity_fac ~ acousticness +
  danceability + energy + instrumentalness + speechiness
  + duration_s + loudness + explicit + year_c_fac +
  year_c_fac:explicit)
```

	Residual Deviance	AIC
model_2	248311 ↓	248371 ↓

Proportional Odds Model 2 Diagnostics

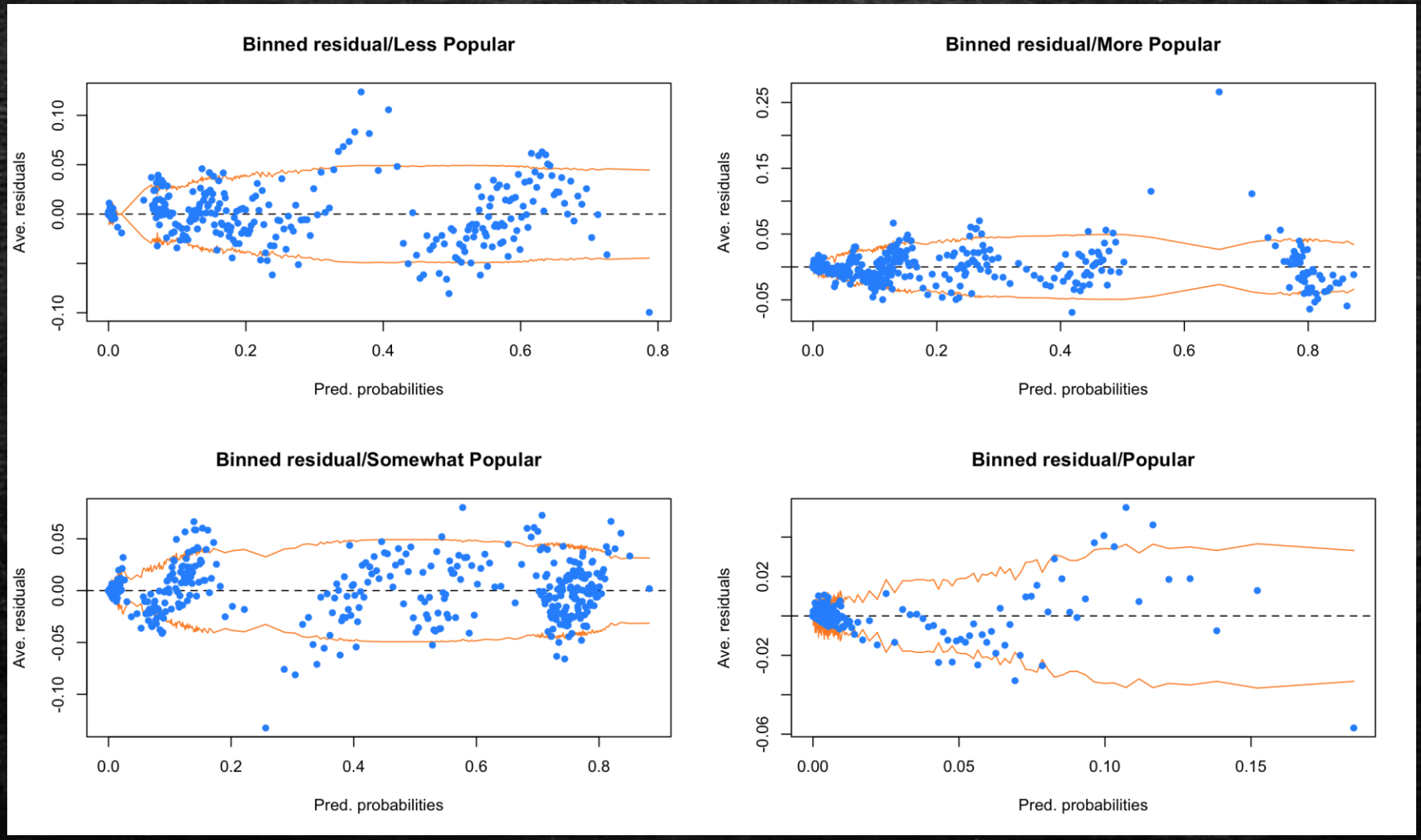


Multinomial Logistic Model

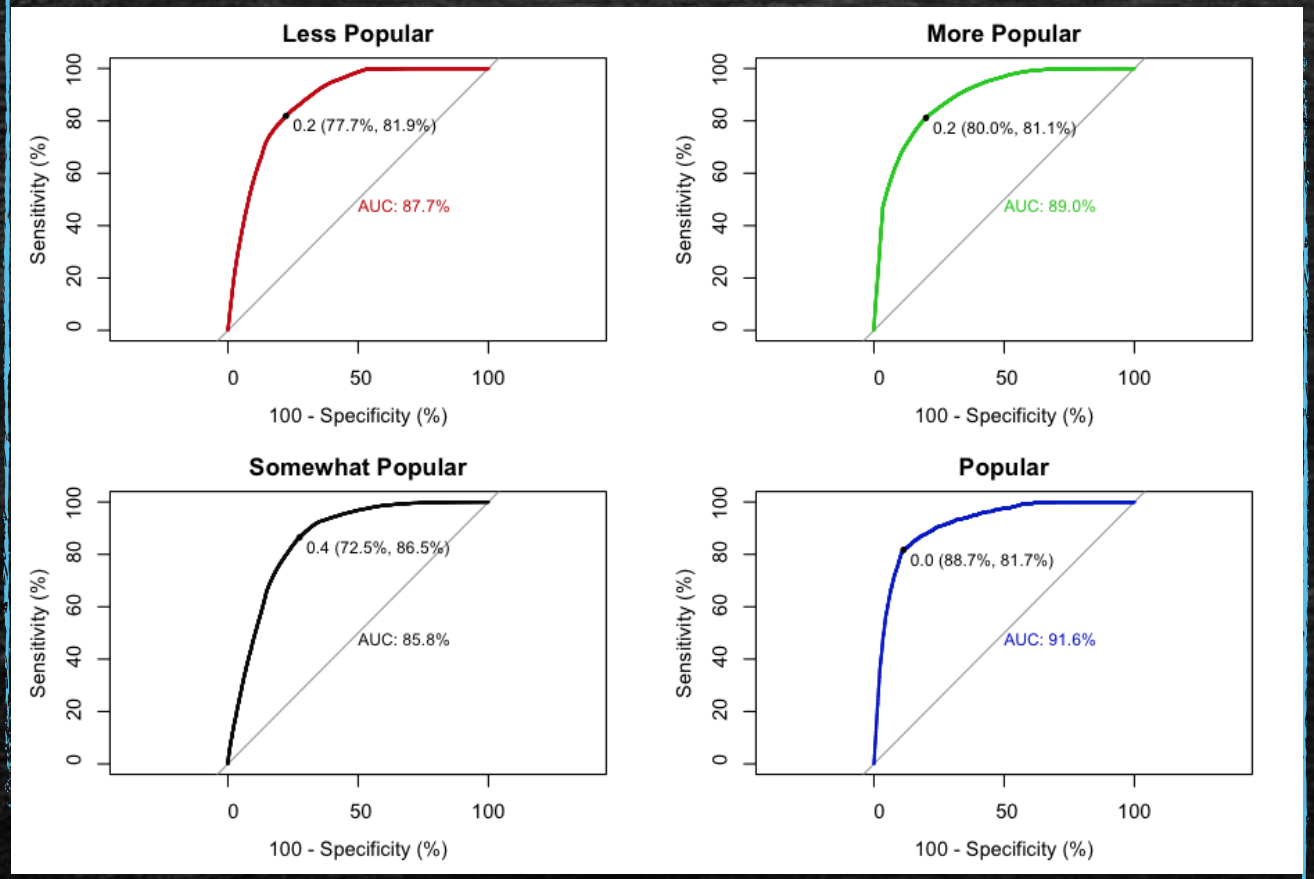
```
model_mlm <- multinom(formula = popularity_fac ~ acousticness
  + danceability + energy + instrumentalness + speechiness +
  duration_s + loudness + explicit + year_c_fac +
  year_c_fac:explicit)
```

	Residual Deviance	**AIC**
model_mlm	239399 ↓	239615 ↓

Multinomial Logistic Model Diagnostics



Accuracy and AUC For All Popularity Levels



Overall Accuracy

0.69

	Sensitivity	Specificity
Class: unrated	0.7817743	0.9487555
Class: less popular	0.6292478	0.8836599
Class: somewhat popular	0.8211817	0.7602092
Class: more popular	0.4804133	0.9629952
Class: popular	0.0000000	1.0000000

Final Model – Multinomial Logistic Model

y.level	term	estimate	std.error	statistic	p.value	conf.low	conf.high
less popular	(Intercept)	2.004900e+00	0.0634	1.097810e+01	0.0000	1.770800e+00	2.270000e+00
less popular	acousticness	9.340000e-02	0.0470	-5.047020e+01	0.0000	8.520000e-02	1.024000e-01
less popular	danceability	3.929000e+00	0.0570	2.399290e+01	0.0000	3.513500e+00	4.393700e+00
less popular	energy	4.242000e-01	0.0638	-1.343960e+01	0.0000	3.743000e-01	4.807000e-01
less popular	instrumentalness	6.802000e-01	0.0305	-1.263740e+01	0.0000	6.407000e-01	7.221000e-01
less popular	speechiness	1.184000e-01	0.0611	-3.494120e+01	0.0000	1.051000e-01	1.335000e-01
less popular	duration_s	9.998000e-01	0.0001	-3.071700e+00	0.0021	9.996000e-01	9.999000e-01
less popular	loudness	9.931000e-01	0.0025	-2.826600e+00	0.0047	9.883000e-01	9.979000e-01
less popular	explicit1	8.900000e-03	0.0700	-6.743900e+01	0.0000	7.800000e-03	1.020000e-02
less popular	year_c_fac1930-1939	1.609900e+00	0.0458	1.038780e+01	0.0000	1.471600e+00	1.761200e+00
less popular	year_c_fac1940-1949	1.184100e+00	0.0437	3.864300e+00	0.0001	1.086800e+00	1.290000e+00
less popular	year_c_fac1950-1959	6.624500e+00	0.0425	4.449990e+01	0.0000	6.095200e+00	7.199800e+00
less popular	year_c_fac1960-1969	8.149913e+08	0.0256	8.012982e+02	0.0000	7.750976e+08	8.569382e+08
less popular	year_c_fac1970-1979	1.115956e+08	0.0395	4.688842e+02	0.0000	1.032779e+08	1.205831e+08
...							
somewhat popular	duration_s	9.998000e-01	0.0001	-1.730500e+00	0.0835	9.996000e-01	1.000000e+00
...							
somewhat popular	year_c_fac1940-1949	1.049800e+00	0.0789	6.160000e-01	0.5379	8.993000e-01	1.225500e+00
...							
more popular	duration_s	1.000000e+00	0.0001	-3.329000e-01	0.7392	9.997000e-01	1.000200e+00
...							

- More observations in the 95% se bound
- Observations have a more random pattern inside the bound
- Almost all variables are significant (104/108)

Conclusion (Popularity Level: Popular)

Positive Influence	Negative Influence
danceability	acousticness
loudness	energy
explicit1	instrumentalness
year_c_fac1950-1959	speechiness
year_c_fac1960-1969	duration_s
year_c_fac1970-1979	year_c_fac1930-1939
year_c_fac1980-1989	year_c_fac1940-1949
year_c_fac1990-1999	
year_c_fac2000-2009	
year_c_fac2010-2020	
explicit1:year_c_fac1930-1939	explicit1:year_c_fac1970-1979
explicit1:year_c_fac1940-1949	explicit1:year_c_fac1990-1999
explicit1:year_c_fac1950-1959	explicit1:year_c_fac2010-2020
explicit1:year_c_fac1960-1969	
explicit1:year_c_fac1980-1989	
explicit1:year_c_fac2000-2009	

Research Questions Recap

- Potential determinants for song popularity on Spotify
- Are certain audio features affect popularity more than others? (if so, to what extent?)
- Are these features different for songs in different popularity levels?
- Any interesting interactions between audio features that affect song popularity?

Potential Limitations

- Imbalanced data
  - Spotify's user base is dominated by Millennials
- Insufficient and biased in determining popularity
  - Spotify's popularity metric
  - Stream count
- Indepth analysis by genre