Search

Home

Library

Regression Analysis
Spotify Dataset

Team HighlandR

Tristan Dull, Daniel Khan, Aaron Bajorunas, Liam Daly, Christopher Chen

Table of Contents

Intro + Analysis Goal

- We'll be analyzing a real-world Spotify dataset with 2400 observations

- Our main analysis goal is to create a linear regression model to make inferences on **track popularity** .

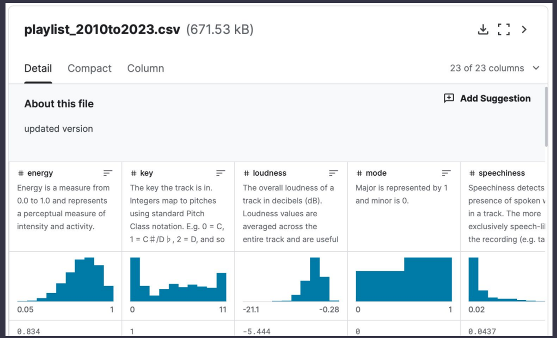
Stakeholders

- Our stakeholders are music industry executives interested in examining which specific factors are linearly related to track popularity.

- Using our results, they may be able to distribute resources toward improving certain factors in order to increase popularity.

Data Description

- From Kaggle.com, extracted from Spotify’s API.
- 2400 observations of 23 distinct variables.
- 100 observations per year from ‘Top Hit’ playlists from 2000-2023.



- **Playlist related:**
  - Playlist url, Year
- **Track related:**
  - Track_id, Track_name, Track_popularity
- **Audio Features:**
  - Danceability, Energy, Key, Loudness, Mode, Speechiness, Acousticness, Instrumentalness, Liveness, Valence, Tempo, Duration_ms, Time_signature
- **Album related:**
  - Album (name)
- **Artist related:**
  - Artist_id, Artist_name, Artist_genre, Artist_popularity

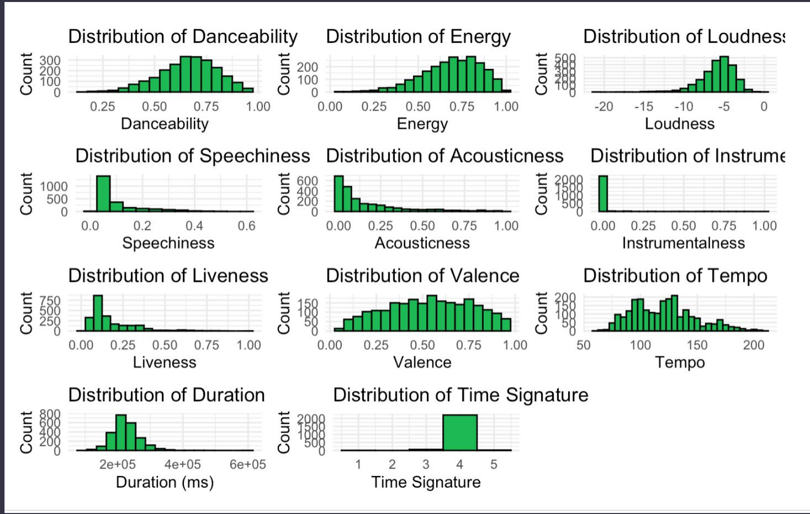
Response Variable: Track_Popularity (double): numerical score of track popularity on a scale from 0 to 100.

Data Cleaning

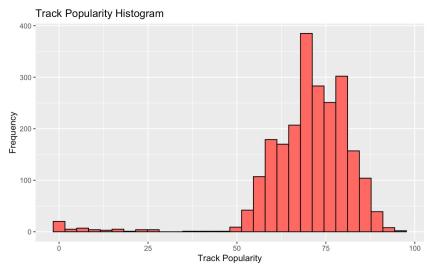
Removed - identifiers
playlist_url
track_id
track_name
album
artist_id
artist_name
artist_genre

Kept	
mode	danceability
speechiness	energy
acousticness	loudness
instrumentalness	year
liveness	key
valence	artist_popularity
tempo	time_signature
duration	

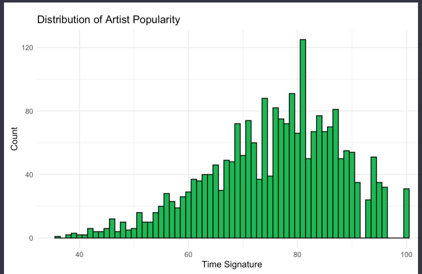
EDA: Variable Distributions



- Danceability, Energy, Loudness, and Time Signature are all left skewed.
- Speechiness, Acousticness, Instrumentaness, Liveness, and Duration are right skewed
- Valence appears approximately normally distributed, Tempo has a fluctuating distribution

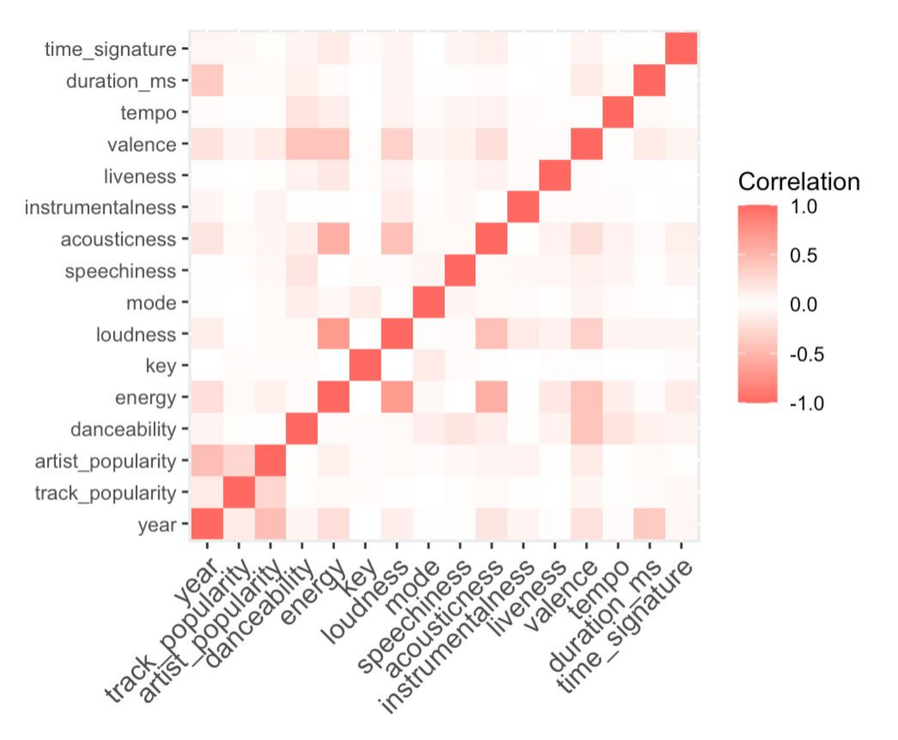


- Track Popularity: Highly left skewed
- Most tracks have a popularity score of 50 or above



Artist popularity appears left skewed

EDA: Correlation Heatmap



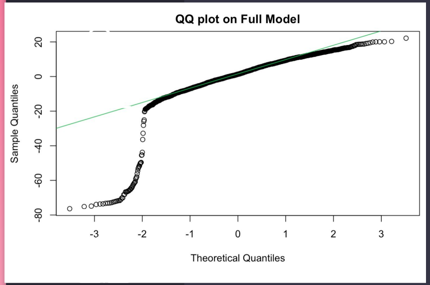
- Heatmap indicates no major concerns for multicollinearity.
- You'll notice somewhat high correlations between the following variables
  - Acousticness and Energy
    - Corr. = -0.55
  - Loudness and Energy
    - Corr. = 0.69
- Both are below the absolute value threshold of .7 so we'll keep them for now.

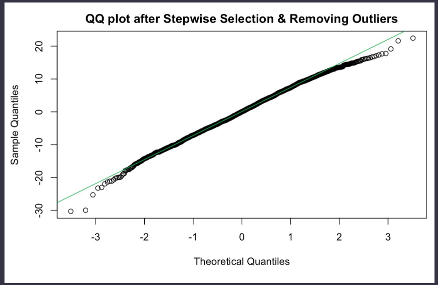
Model Building: Full Model

```
Call:
lm(formula = track_popularity ~ year + artist_popularity + danceability +
    energy + key + loudness + mode + speechiness + acousticness +
    instrumentalness + liveness + valence + tempo + duration_ms +
    time_signature, data = spotify)

Residuals:
    Min      1Q  Median      3Q     Max
-76.361  -4.055   1.326   7.111  22.151

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)        2.013e+01  9.716e+01   0.207   0.8359
year               1.940e-02  4.826e-02   0.402   0.6878
artist_popularity  2.903e-01  2.466e-02  11.773   <2e-16 ***
danceability      -2.723e-01  2.255e+00  -0.121   0.9039
energy             7.813e-01  2.598e+00   0.301   0.7636
key               -8.418e-02  7.327e-02  -1.149   0.2507
loudness           5.649e-02  1.784e-01   0.317   0.7516
mode               3.036e-02  5.400e-01   0.056   0.9552
speechiness       -3.911e+00  2.886e+00  -1.355   0.1754
acousticness       1.220e+00  1.532e+00   0.796   0.4259
instrumentalness   1.341e+00  3.219e+00   0.417   0.6770
liveness           3.389e-01  1.974e+00   0.172   0.8637
valence           -1.338e+00  1.450e+00  -0.923   0.3563
tempo             -7.565e-03  9.770e-03  -0.774   0.4388
duration_ms       -5.967e-06  7.010e-06  -0.851   0.3948
time_signature    -1.918e+00  1.091e+00  -1.758   0.0789 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 12.5 on 2286 degrees of freedom
Multiple R-squared:  0.08181,	Adjusted R-squared:  0.07579
F-statistic: 13.58 on 15 and 2286 DF,  p-value: < 2.2e-16
```

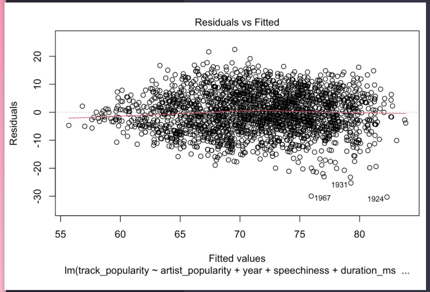
- **Explaining Variation**
  - R squared 0.08181
  - Adj. R Squared = .07579
- 8.181% of variance in track popularity can be explained by our model
- **Significant Predictors**   (a = .10)
  - Artist Popularity
  - Time Signature
- Most significant predictor is artist popularity since it has the highest t value and lowest p value.

Model Building: Assumptions + Stepwise Selection





- Full model fails the normality assumption
- We decided to remove outliers from track_popularity and we used stepwise selection to reduce our model.




```
lag Autocorrelation D-W Statistic p-value
  1       0.6183224     0.7614197       0
Alternative hypothesis: rho != 0
```

- Stepwise identified a total of five significant predictors.
- After reducing our model, normality **substantially improved**

Reduced Model:

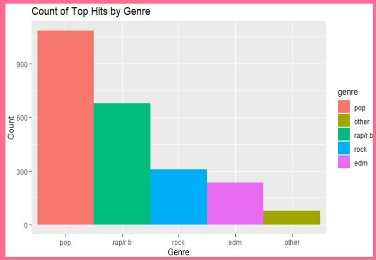
- Constant variance and linearity are **satisfied** .
- Independence assumption is **violated** since DW test indicates autocorrelation.

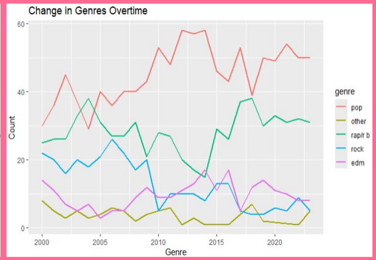
Final Model

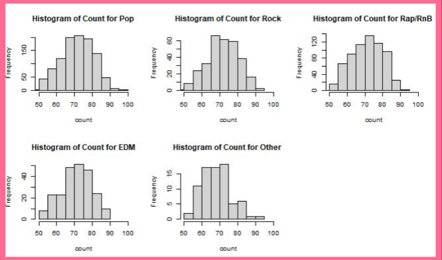
```
Call:
lm(formula = track_popularity ~ artist_popularity + year + speechiness +
    duration_ms + acousticness, data = spotify)

Residuals:
     Min       1Q   Median       3Q      Max
-30.2886  -4.8244   0.0135   5.0460  22.4378

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)       -6.415e+02  5.469e+01 -11.729  < 2e-16 ***
artist_popularity  2.810e-01  1.427e-02  19.685  < 2e-16 ***
year               3.454e-01  2.727e-02  12.665  < 2e-16 ***
speechiness       -6.459e+00  1.626e+00  -3.973 7.32e-05 ***
duration_ms       -1.149e-05  3.971e-06  -2.895  0.00383 **
acousticness       1.763e+00  7.403e-01   2.382  0.01730 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 7.217 on 2240 degrees of freedom
Multiple R-squared:  0.3358,    Adjusted R-squared:  0.3343
F-statistic: 226.5 on 5 and 2240 DF,  p-value: < 2.2e-16
```

- **Explaining Variation**
  - R squared 0.3358
  - Adj. R Squared = .3343
- 33.58% of variance in track popularity can be explained by our model
- **Significant Predictors**
  - Artist Popularity
  - Year
  - Speechiness
  - Duration
  - Acousticness
- Most significant predictor is artist popularity since it’s associated with the largest t value and lowest p value.
- F statistic shows all coefficients are statistically significant

Other Findings

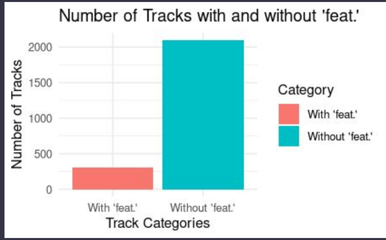


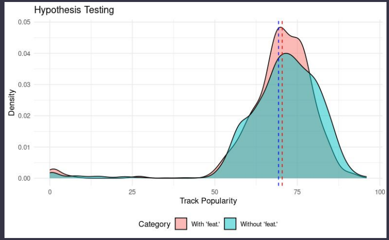




Genres

- Overall there were about 241 genres and some artists were assigned different 11 genres, this was condensed into 5 categories by each artist's foremost genre.
- Pop was the most popular genre by far and has been the most popular over the decades.





$$H_0 : \mu_{feat.} = \mu_{nofeat.}$$
$$H_a : \mu_{feat.} \neq \mu_{nofeat.}$$

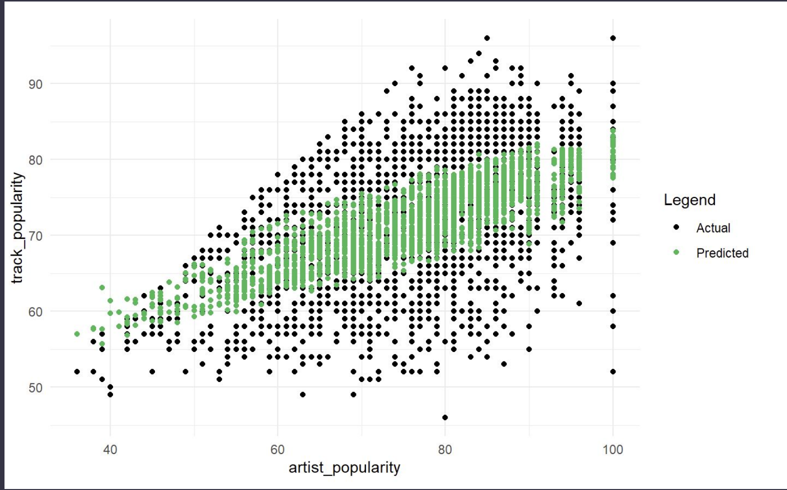
Features

- When determining if there's a difference between popularity of music with and without features, our findings indicate that a song's feature doesn't amplify's a musician's popularity.

Conclusion and Limitations

Limitations:

- Two out of four linear regression assumptions violated
- Shapiro-Wilk shows the data is not normally distributed
- Durbin Watson test shows some autocorrelation (between year and artist popularity)
- low R-squared value of .3358



Bottom Line : Music industry executives should focus on reducing speechiness and duration, as they negatively impact track popularity. They should attempt to work with artists who are already popular in order to maximize track popularity.

Contributions

Tristan D. Organized project direction, stepwise regression, initialized final model, plot of actual values with predicted

Liam D. Organized slides template, cleaned data, created variable visualizations/correlation heatmap, initialized full model.

Chris C. Organized the look of slides, cleaned up genres in the data, created visualizations for EDA and for genre analysis.

Daniel K. Filtered music for with and without features, visualized average song popularity and performed hypothesis testing

Aaron B. Conducted analysis on song duration, danceability, and their relationship with popularity, and creating visualizations and CV