Spotify

Andrew Nguyen
Anne Cuzeau
Samuel Flusche

Welcome

Play Follow

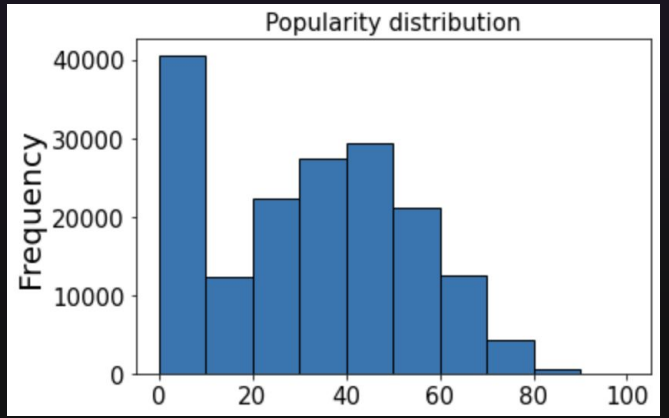
About the data:

- No missing value, number of rows: 170,653
- Includes songs from 1921 - 2021
- Most columns are descriptive variables about the songs
- Discarded columns: Artist name, release date (inconsistent & redundant), Tempo, Speechiness, mode, key

Data exploration

Initial findings:

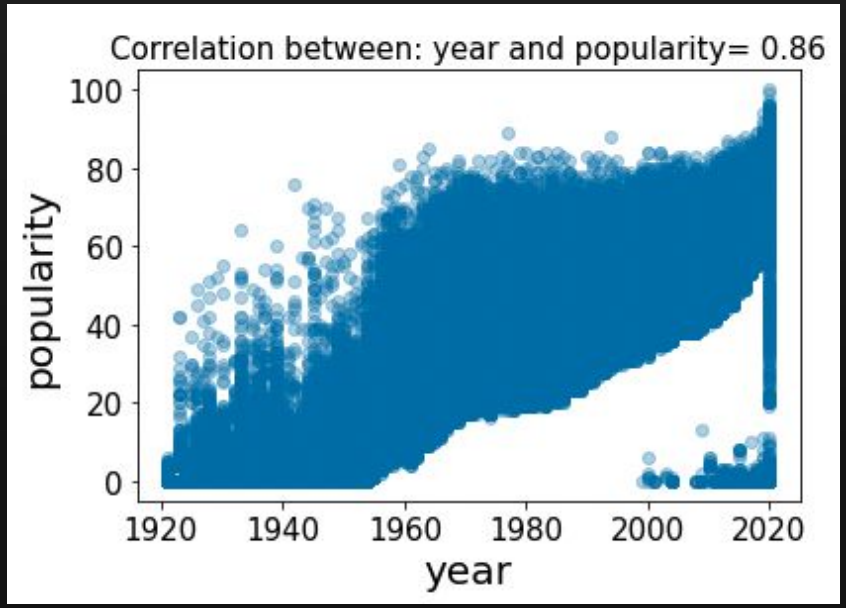
- Upward trend for many factors: loudness, energy, danceability, popularity
- Uneven distribution of popularity:
  - A lot of songs (about 16%) have a popularity of zero, and popular songs are rare
- As we'll see, we dropped some columns that did not make sense (key, mode...)



Year and Popularity

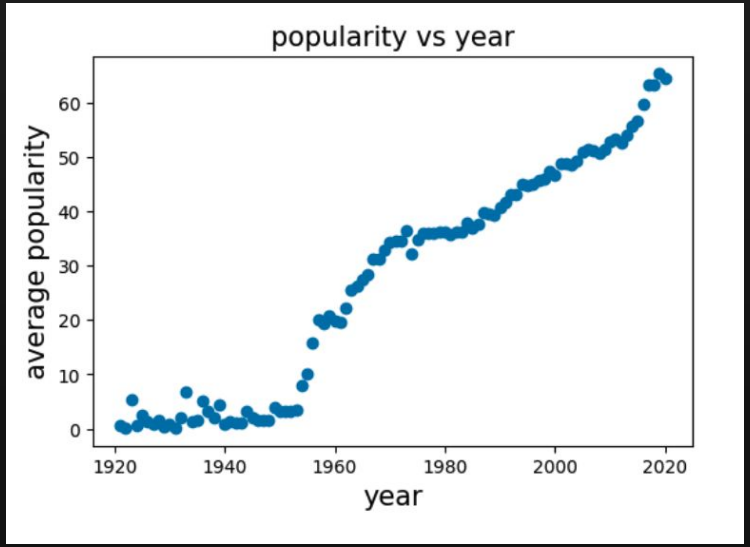
Most linear relationship: Year and Popularity (1-100)

- The most linear relationship: r^2 = 0.74
- Most (very) popular songs are new
- Can probably be explained by Spotify's audience



Year and Popularity

Year and Popularity: Upward trend

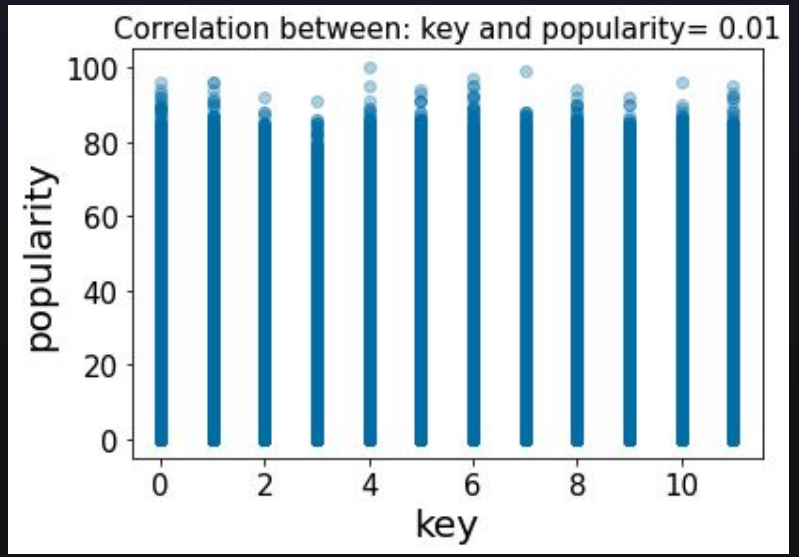


- Average popularity per year goes up
- Pitfall: how can our model predict an old but popular song accurately?

Not Always a Trend

Multiply the Variables:

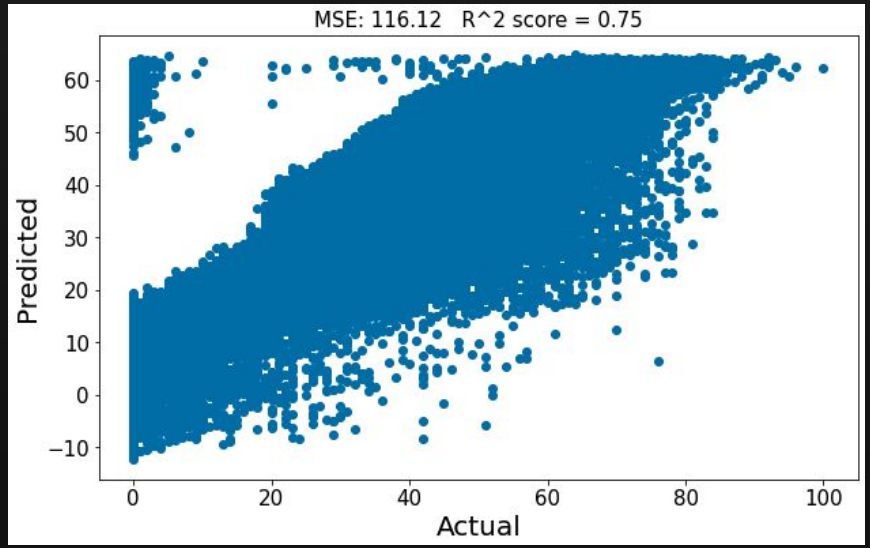
- Some graphs are like this one, with no leading trend
- We dropped these variables:
  - Tempo, Speechiness, mode, key, valence



Baseline Linear Model

Everything, all at once:

- Took all the variables that we did not drop and ran a linear regression
- Split train/test: 70/30
- Max of 60: our model is incapable of predicting any 'popular' song



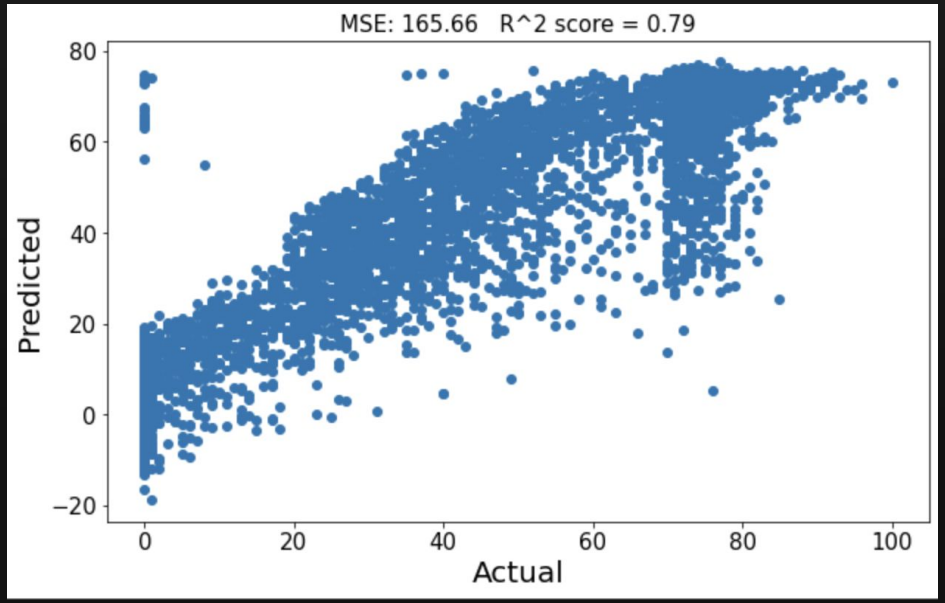
Let's improve our model

Balancing data

- Goal: getting our model to predict popular song
- Methodology:
  - Took the number of songs with popularity >= 70
  - Randomly selected the same amount of songs with popularity 70-30 and 30-0

Let's improve our model

Balancing data



Results:

- Better job (more linear, predicts to 80) except around 70
- MSE is up because we are making 'costly' mistakes between 70 & 80 and below 0.

Let's improve our model

Acousticness: at first, looks like lower acousticness is better!

- Acousticness is actually trending down
- However the relationship between popularity and acousticness changes over time:
  - older songs with low values are more likely to be popular
  - newer songs with higher values are more likely to be popular

Let's improve our model

Other factors

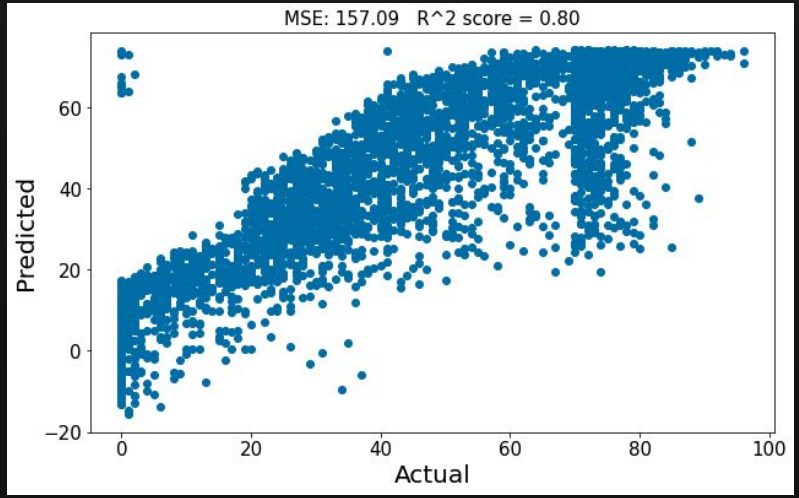
- Energy and loudness seemed like good candidates but after 'de-trending': popular and unpopular songs have the same range regardless of the year
- Instrumentalness: trending down, but lower is better across years
- Duration & Danceability: popular and unpopular songs have the same range regardless of the year
- Look for relationships between factors: only loudness and energy were correlated (0.78)

Putting it all together

Calculated coefficients, year and balanced data

Results:

- Still not predicting to 100
- Higher r^2: slightly better fit
- MSE is still up (same mistakes) between 70 & 80 and below 0.

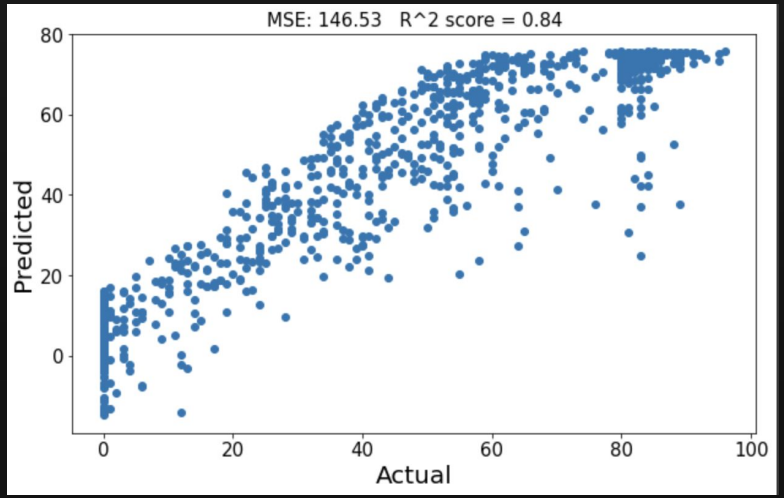


Putting it all together

Narrower samples: popularity >80, 79-50, 49-20, <20

Results:

- As expected: smaller training set
- Better r^2: better fit, up by 9 points
- MSE is down: fewer mistakes but smaller data set
- Still not predicting 100

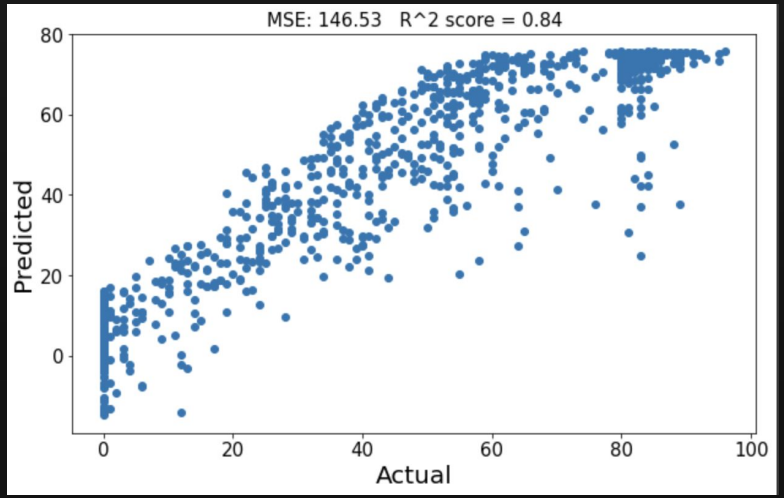


Wrapping Up

Our best model

Why it is our best:

- Lower MSE, highest r^2 → more linear relationship
- Not as many mistakes for high values
- Still not predicting extreme (low and high) values accurately



Wrapping Up

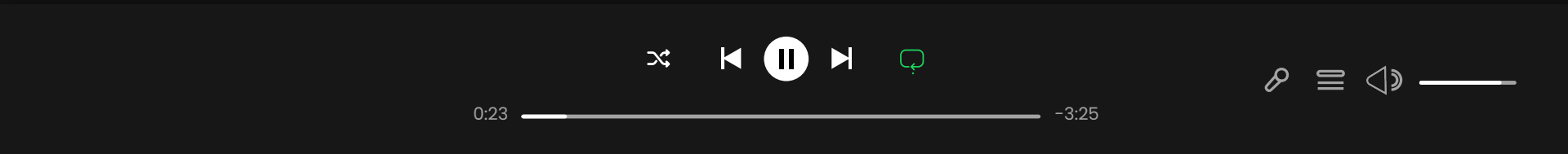
Our model's most 'costly' mistakes

Methodology: Top 20 songs with the largest difference between predicted and actual popularity

→ Almost all songs were unpopular songs (0) predicted to be between 65-80

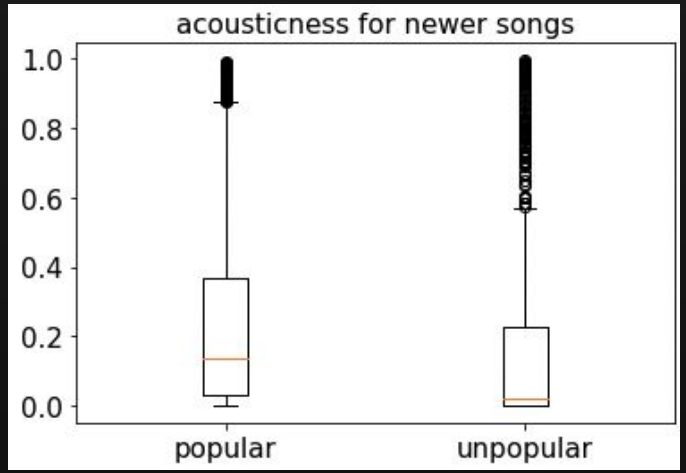
EXCEPT for "White Christmas": 78 in popularity, 5 according to our model

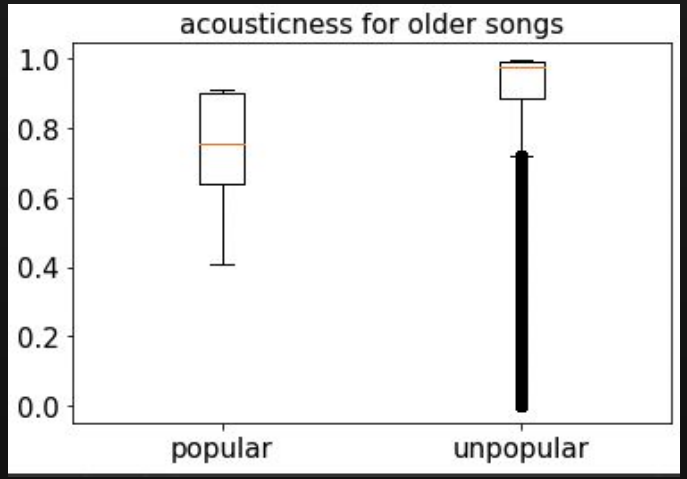
Thank You



Acousticness for older & newer songs

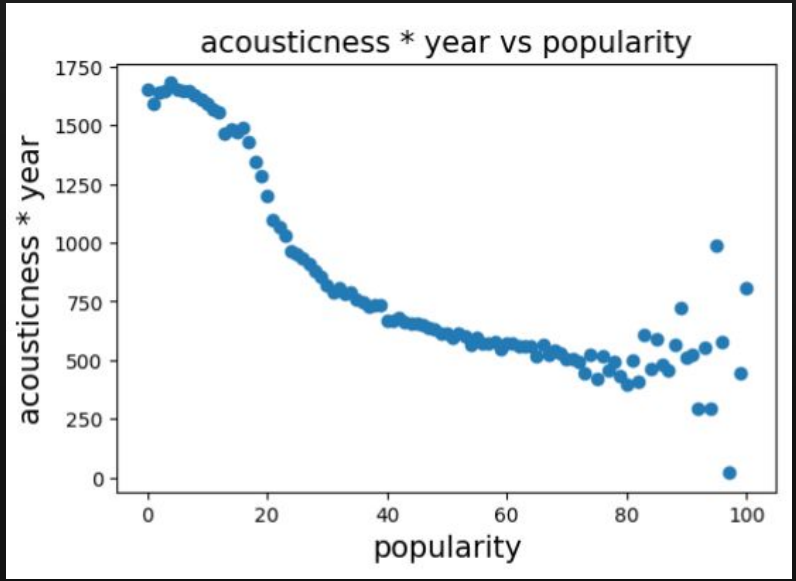
Newer: 2000-2021, Older: <1960





Further Assessing the Relationship

Multiply the Variables:



- Multiplying columns shows relationship with no extraneous variables