First Week Streams Predictor

[Anthony J. Rumph]

The challenge

Predict Number of Streams First Week

Create a tool that allows anyone to predict the amount of streams a given song will receive during its first week listed with Spotify.

THE APPROACH

We use a mix of techniques including:

- API Connections
- Web Scraping
- Natural Language Processing
- Logistic Regression
- Model Evaluation

SUCCESS

Build an application that predicts stream count and provides impact weights for each feature

Raw Data Details

- 35 Weeks of data
- 1,184 Songs
- 43 Features

For Each Song Gather:

- Song Information - Spotify
- Artist Information - Spotify
- Song Audio Analysis - Spotify A*
- Song Lyrics - Genius
- Tweets - Twitter

Data Gathering & Formatting Process

Gathering Our Data:

- Spotify Song Search API - Basic info: duration, title, artist...
- Spotify Artist Search API - Artist followers, popularity, genres, albums...
- Genius API - Song versions and lyrics
- Twitter API - User's tweets about artist

Data Transformation Included:

- Standardize track name formats
- Remove duplicates from chart re-entry
- Remove songs without corresponding lyrics

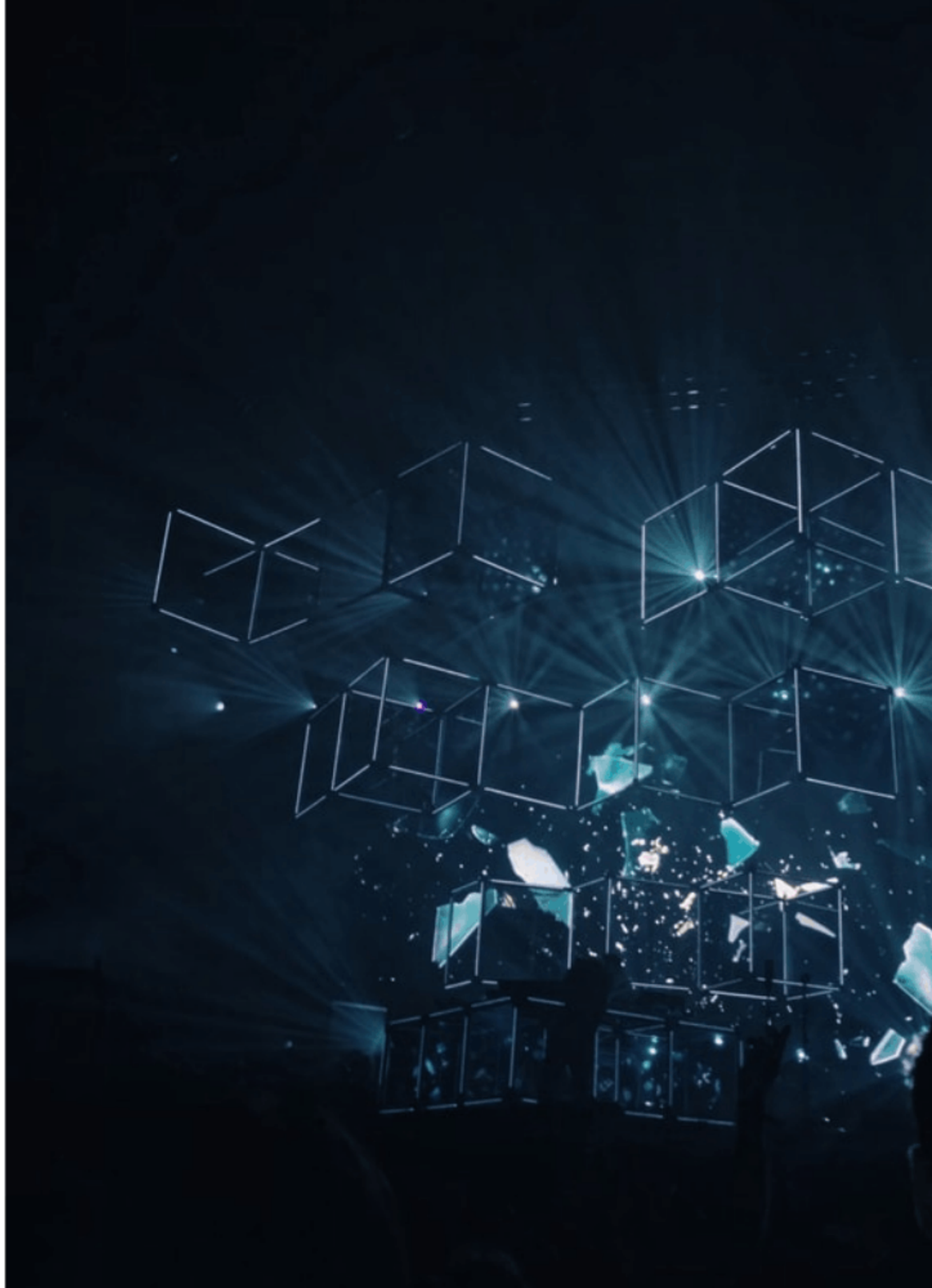
Exploratory Data Analysis

Key Explorations

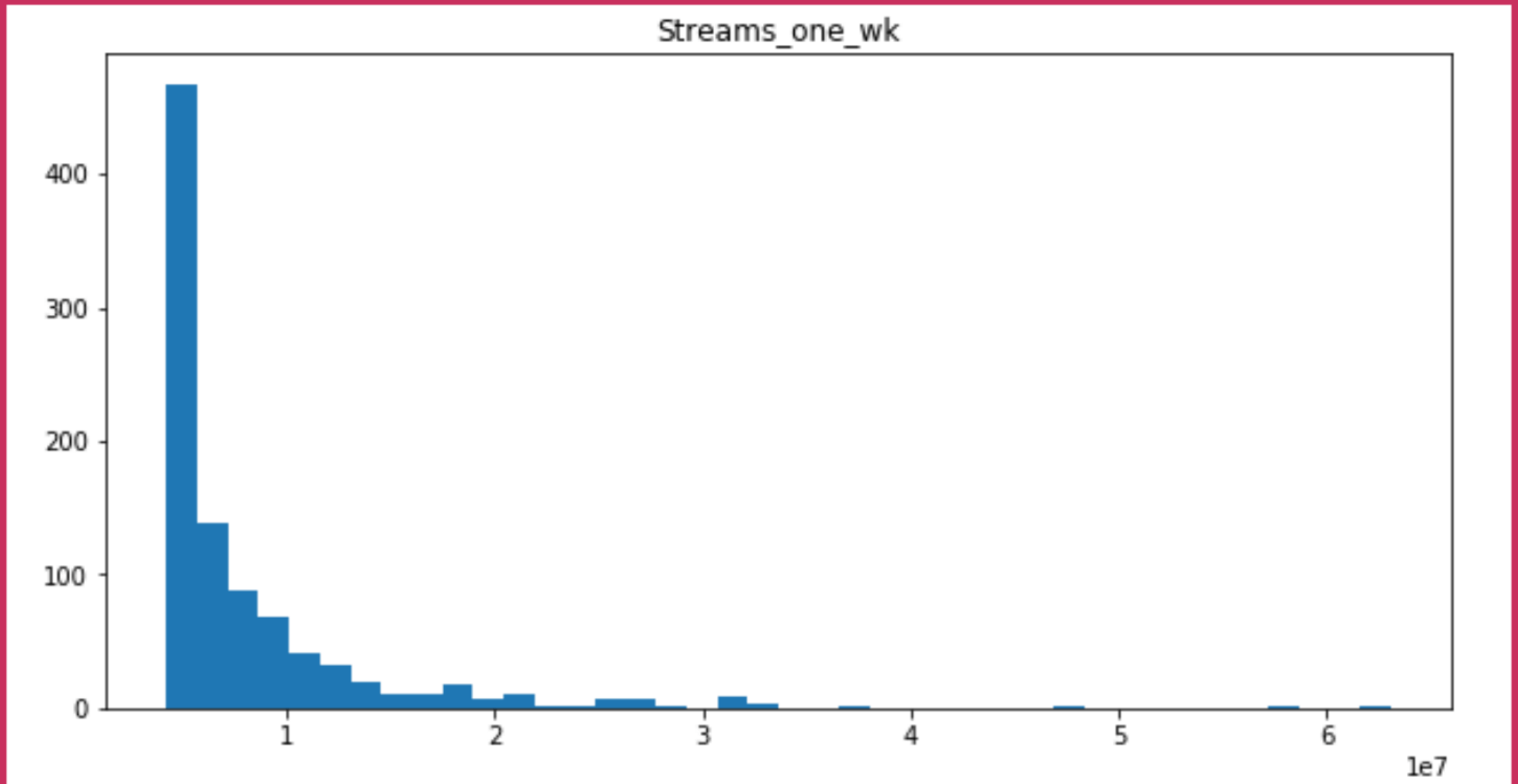
Let's examine:

1. Stream Distributions
2. Followers
3. Popularity
4. Data Completeness
5. Lyric Sentiment Distribution

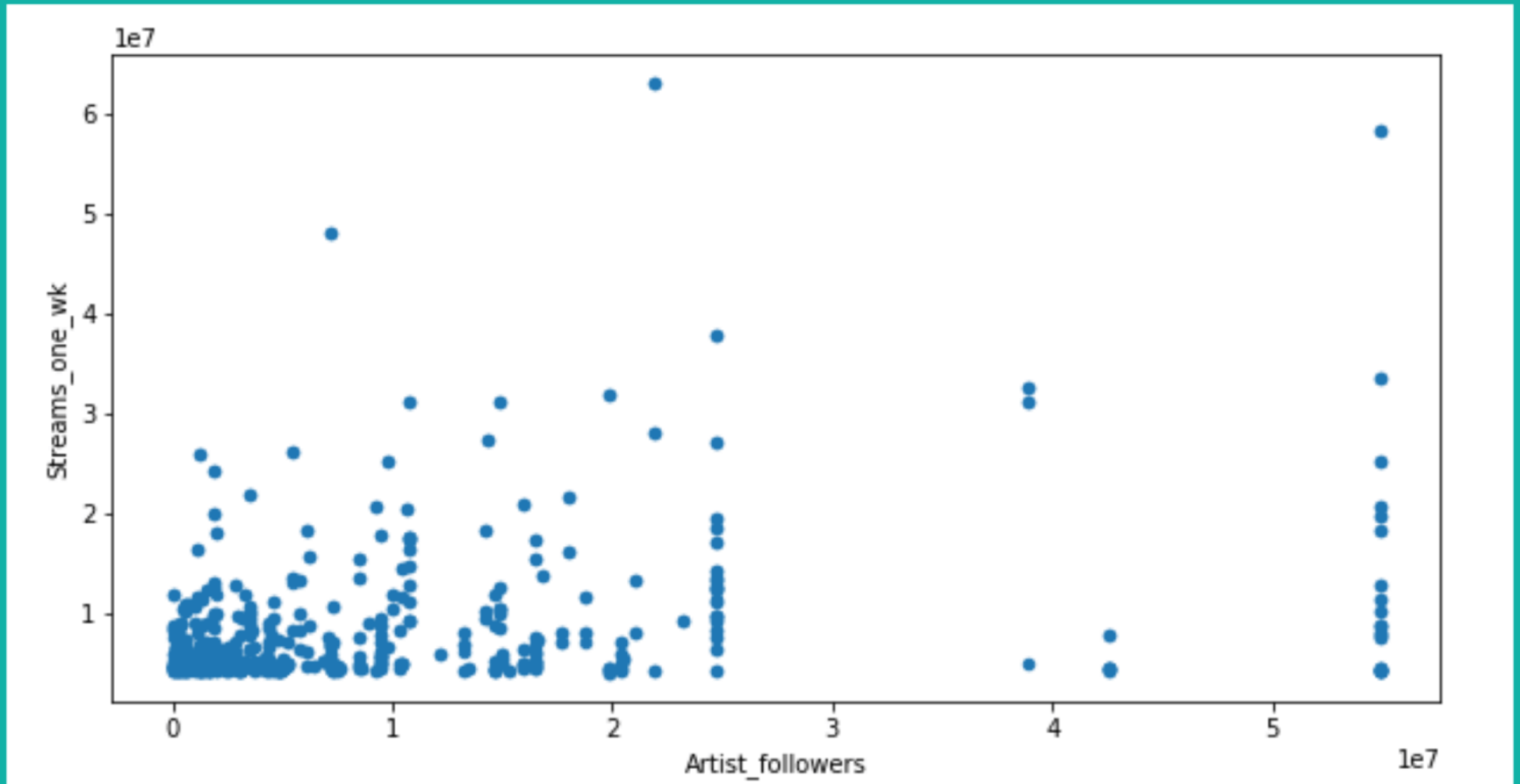
As experienced with our current dataset



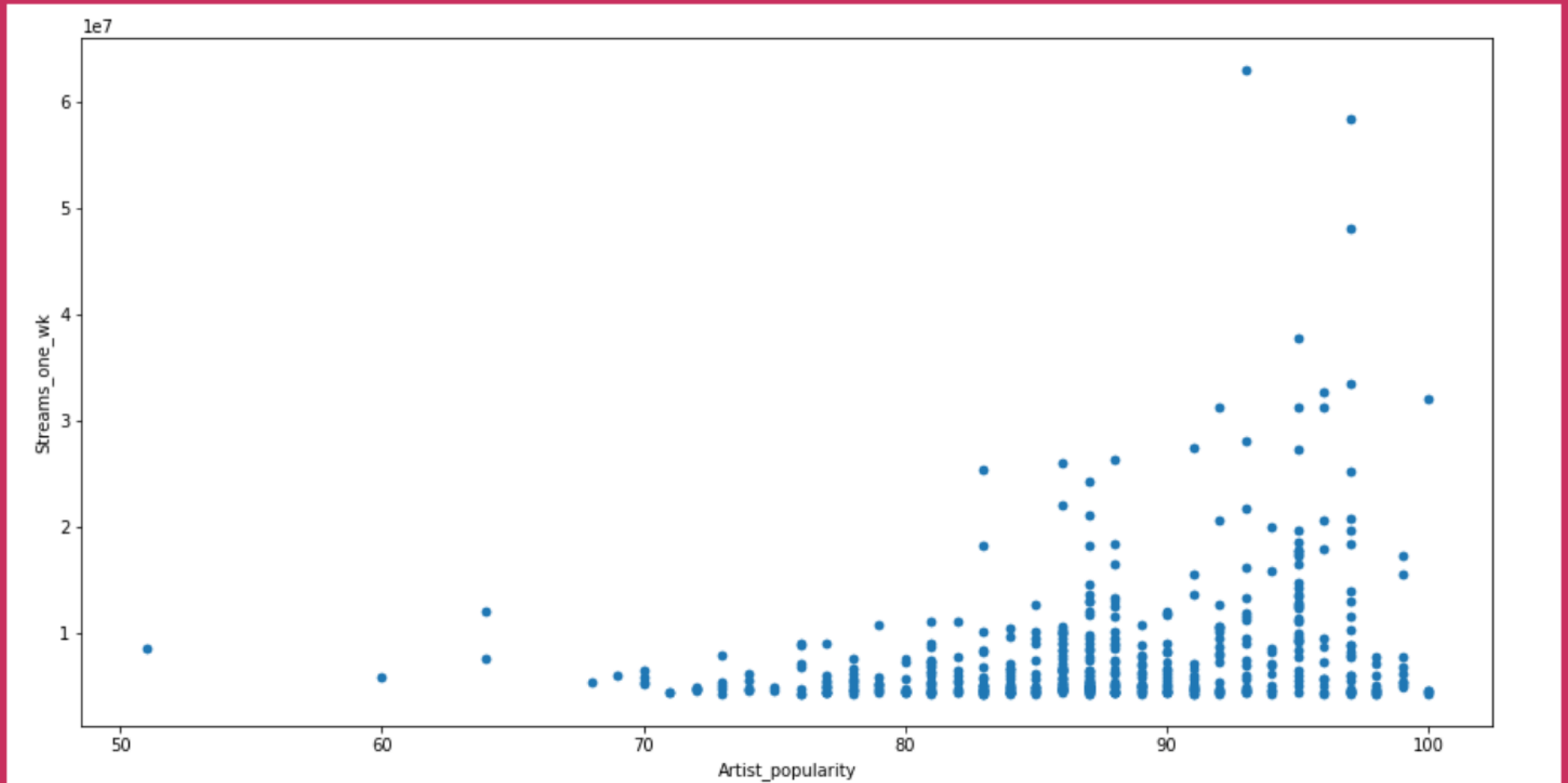
Left hand skewed curve as majority of songs debut with less than 1 million streams first week.



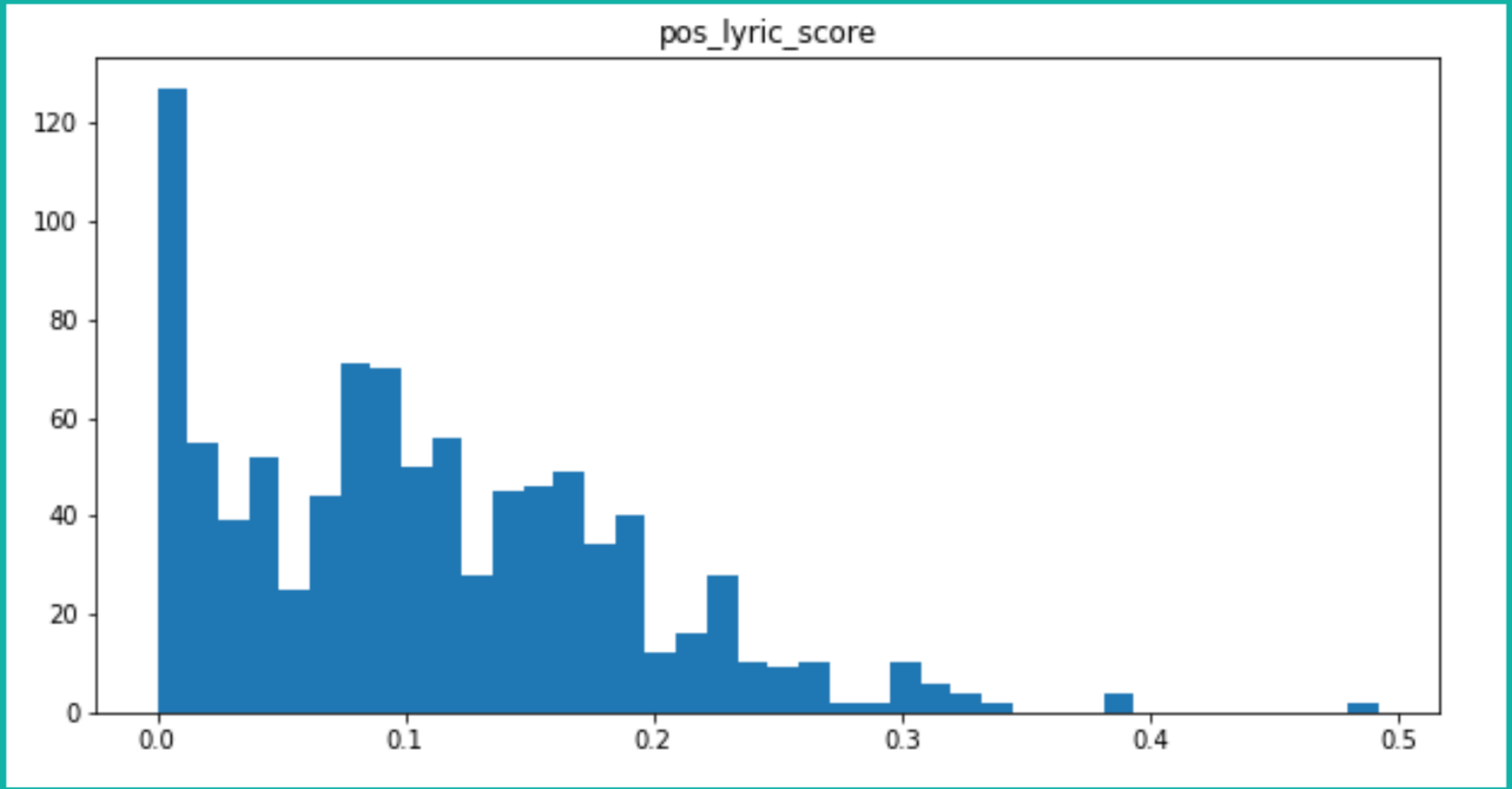
Not much of a direct correlation. Several artist with many followers who received low stream counts.



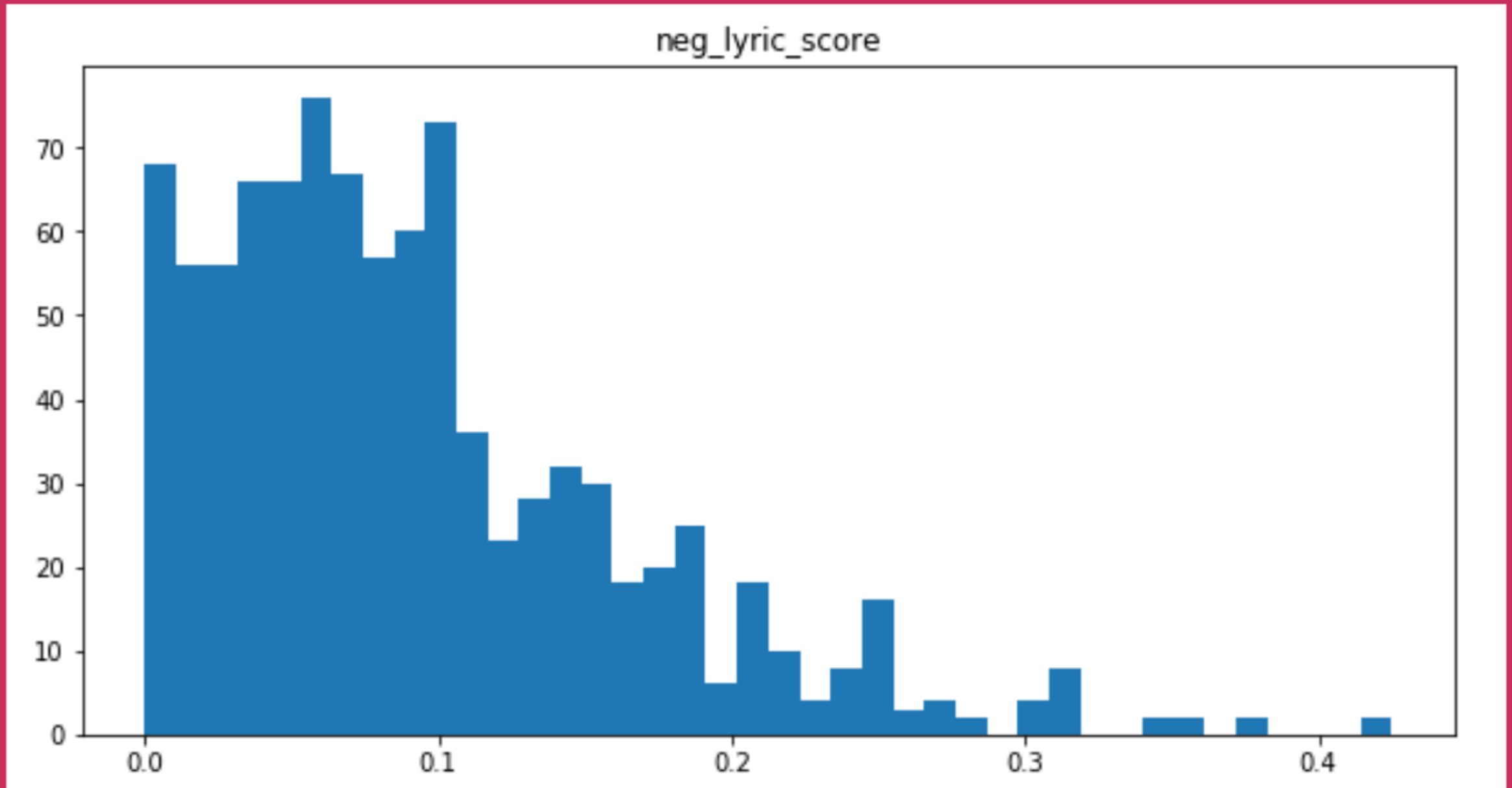
More of the expected direct relationship.
Greater popularity yields greater streams.



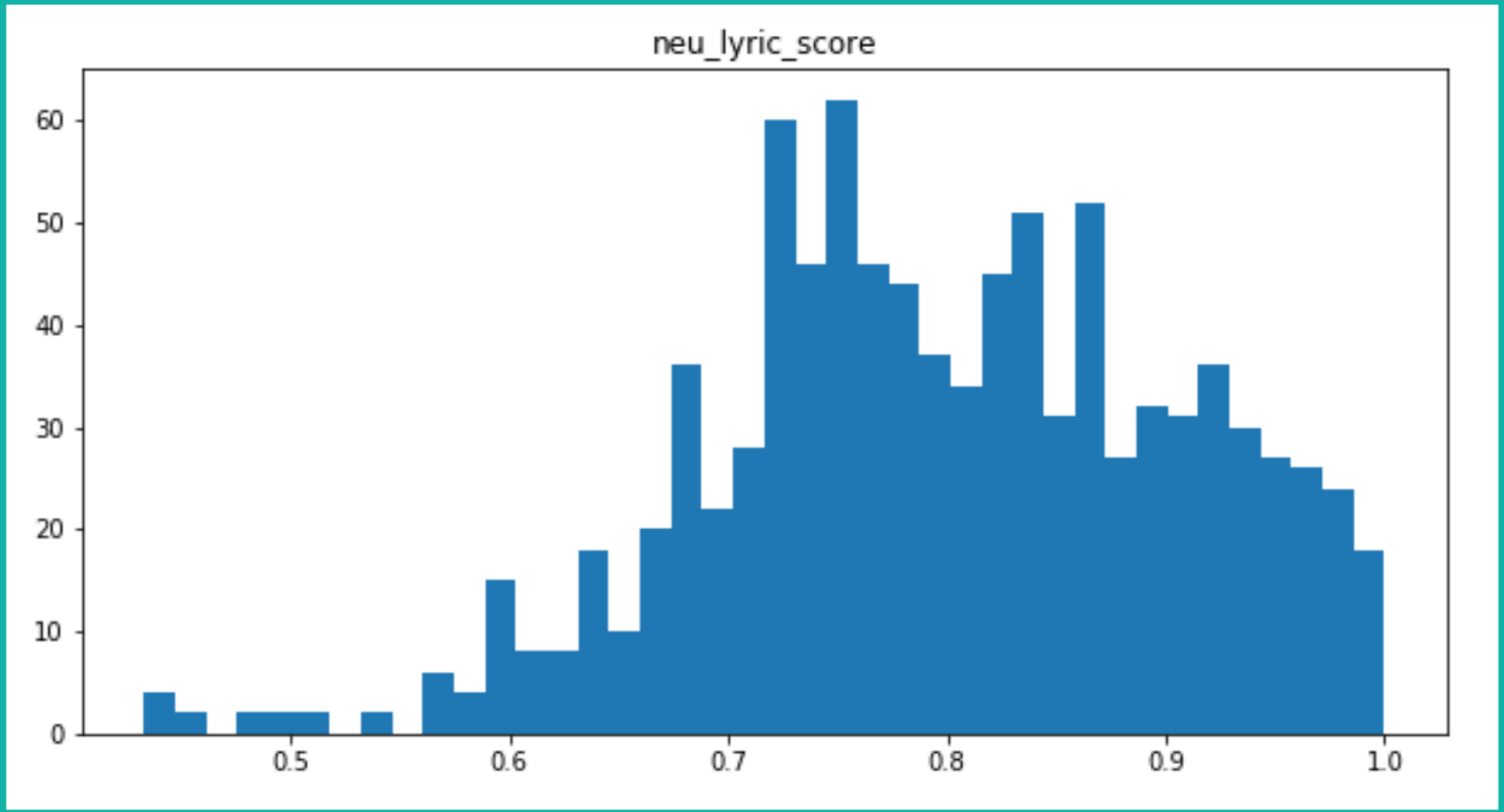
Most tend toward 0 or "less positive".



Hudled toward 0 or "less negative" yet not with the same intensity as the positive.



More tend toward 0 which as expected.



Prediction Model

Linear Regression Model

Primary Model Details:

- 80 / 20 Train-Test Split
- All Numerical Features
- Avg R2 Score of 23%

Most Impactful Features:

Lyric Scores are actually the strongest positive signals

Acousticness and Danceability are the strongest negative indicators.

14	2.247465e+08	pos_lyric_score
13	2.210594e+08	neu_lyric_score
12	2.139679e+08	neg_lyric_score
9	1.248370e+06	valence
8	9.244981e+05	liveness
15	5.567955e+05	com_lyric_score
21	1.592458e+05	Artist_popularity
3	5.834846e+04	loudness
2	3.964361e+04	key
20	7.297673e-02	Artist_followers
19	0.000000e+00	com_twitter_artist_score
18	0.000000e+00	pos_twitter_artist_score
17	0.000000e+00	neu_twitter_artist_score
16	-2.384186e-07	neg_twitter_artist_score
11	-1.165475e+01	duration_ms
10	-8.422243e+03	tempo
1	-4.692857e+05	energy
4	-8.573822e+05	mode
5	-1.055383e+06	speechiness
7	-2.849657e+06	instrumentalness
0	-3.877599e+06	danceability
6	-3.992321e+06	acousticness

Conclusions

- Its not simple to predict a songs first weeks streams based on the data points provided.
- Currently Lyric scores are our strongest signal
- The model has not good way to understand the music
- I need to continue to strengthen this model.

Thank You

Anthony Rumph