

Summary: Spotify and YouTube song metrics and their influence on song performance

Author: Ignacy Smoliński

Dataset introduction

Metrics summary

Dataset consists of both categorical and numerical metrics, with this research focusing on the following:

- Numeric:
 - 'Danceability',
 - 'Energy',
 - 'Key',
 - 'Loudness',
 - 'Speechiness',
 - 'Acousticness',
 - 'Instrumentalness',
 - 'Liveness',
 - 'Valence',
 - 'Tempo',
 - 'Duration_ms',
 - 'Views',
 - 'Likes',
 - 'Comments',
 - 'Stream'
- Categorical:
 - 'Album_type',
 - 'Licensed',
 - 'official_video'

The following columns have been removed from the dataset due to the limited information they provide:

- 'Description',
- 'Url_youtube',
- 'Url_spotify',
- 'Uri',
- 'Title',
- 'Channel',
- 'Album',
- 'Track'

The following metrics have been normalized to 0...1 ranges to avoid negative values:

- 'Loudness'

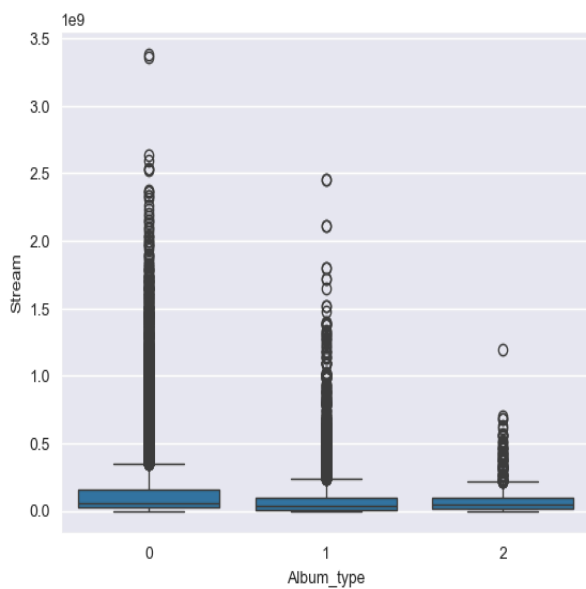
The following categorical metrics have been encoded:

- 'Album_type':
 - Album → 0 (most common value)
 - Single → 1 (the second most common value)
 - Compilation → 2 (rarely seen value)

All rows with missing values have been dropped (less than 2.5% of the dataset in total)

Removing outliers

Due to extremely large differences in values for numeric columns such as 'Stream', 'Views' or 'Likes', with some differences reaching even six magnitudes (detailed data in numeric_summary.csv file) as we compare both global hits and niche songs by fledgling singers, it is crucial to eliminate the outliers before proceeding further. Sample boxplot without having removed the outliers below:



As such, using the IQR method with 1.5 scale to clip outliers for numeric columns and to drop those with more than 20% outliers (too unreliable data). Columns dropped:

- 'Instrumentalness' → 21.34% of the dataset

Despite having clipped the biggest outliers, they still overall retain a strong influence over the dataset. For that very reason, one correlation heatmap will be done by comparing the best performing singers only.

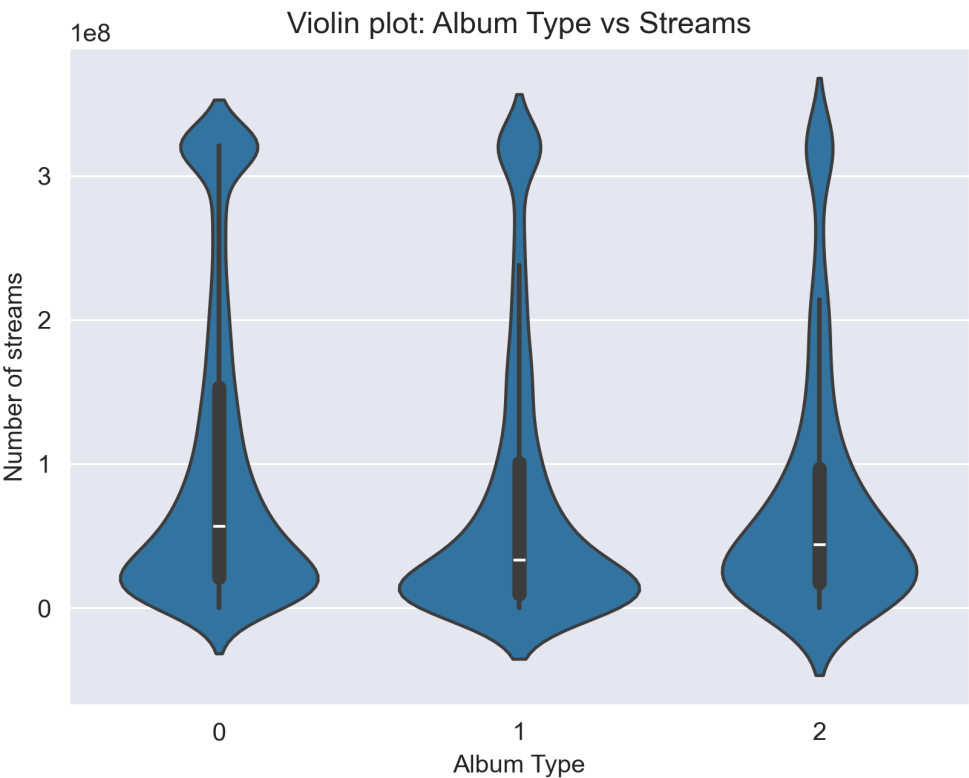
Statistics from the clipped dataset

Researching the clipped dataset allows us to see how various metrics influence the songs' performance as a whole. The most interesting of observations described in detail below.

Album type as the deciding factor for most listened to songs

More listened to artists tend to publish their songs in sets as albums. According to the researched on data, such an approach tends to increase the average number of streams (times a song was listened to on Spotify).

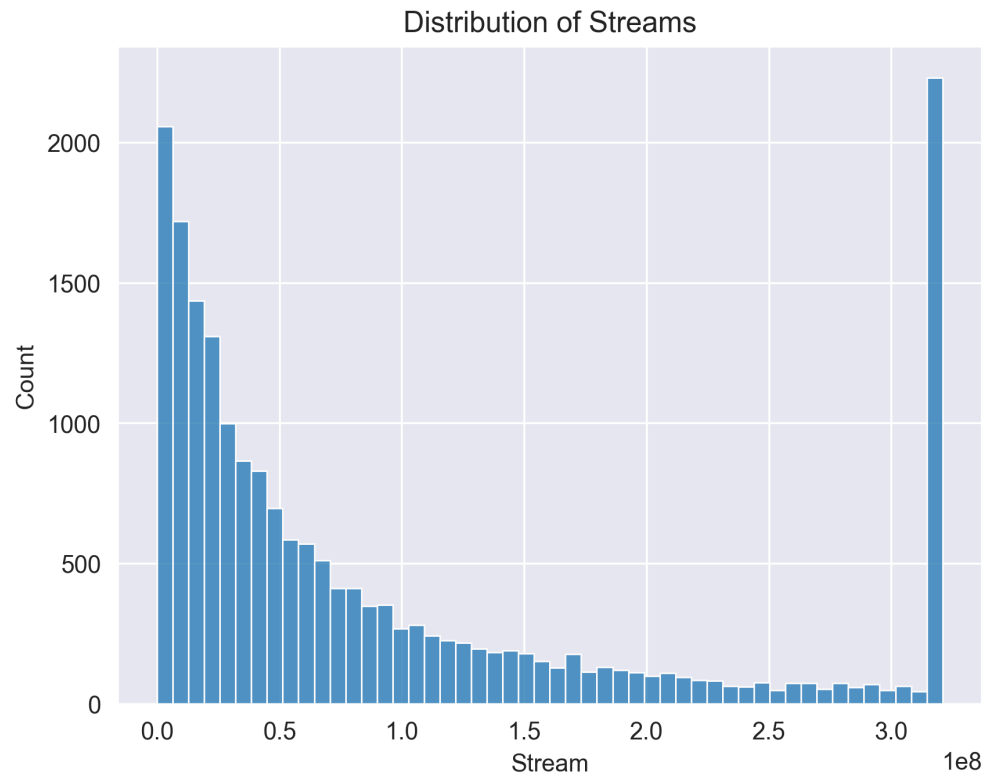
Presented as a violin plot:



When looking at the distribution (width) of the violin plot, songs released as a single appear to be far less dominant at the top — higher stream count. The histograms below provide additional information into the distribution.

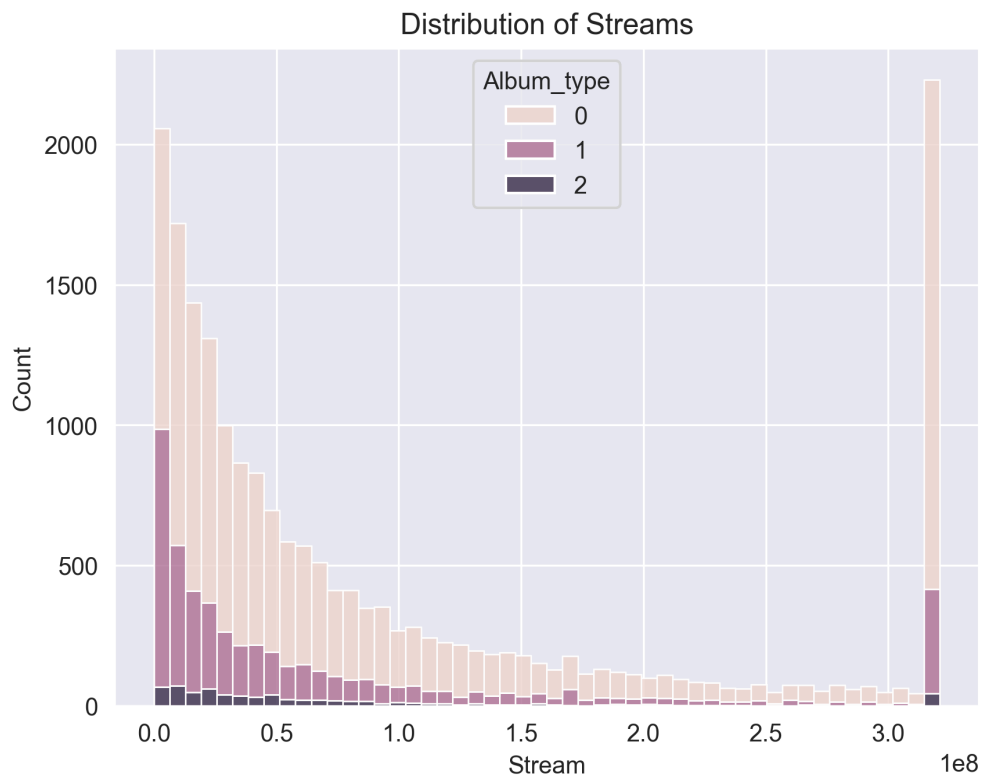
Presented as a histogram with hue added:

- default



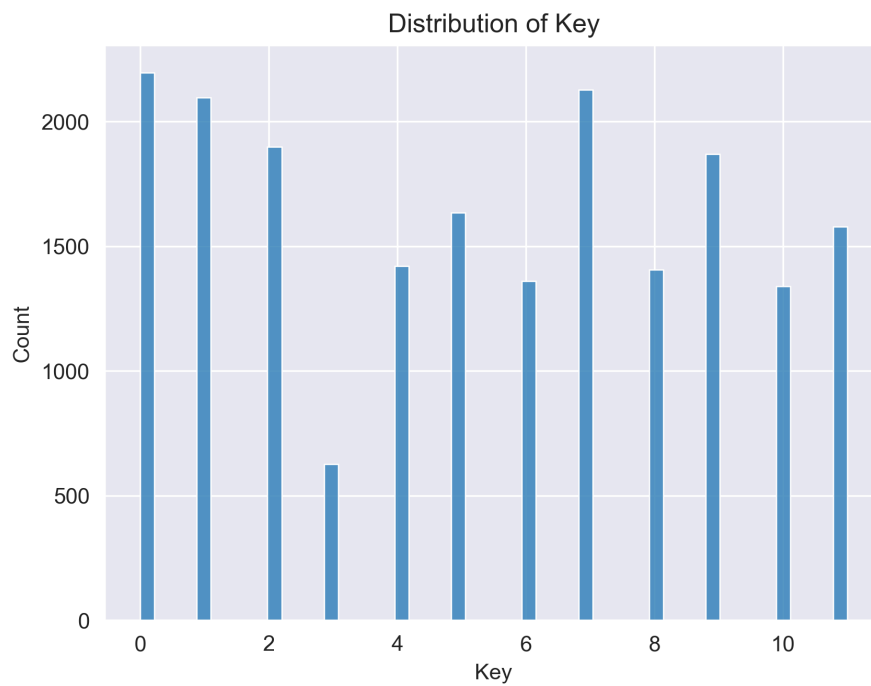
The first histogram of stream count with album type hue added to it shows the difference more clearly, especially in the last, highest bucket. However, as more songs are released as a part of an album, comparing the counts directly may not be the best idea. As such, an independently normalized histogram below might provide more reliable insight.

- independently normalized for density



Judging from it, it can be seen that singles are more common in the lower ranges and albums in higher ranges which would prove the conjecture.

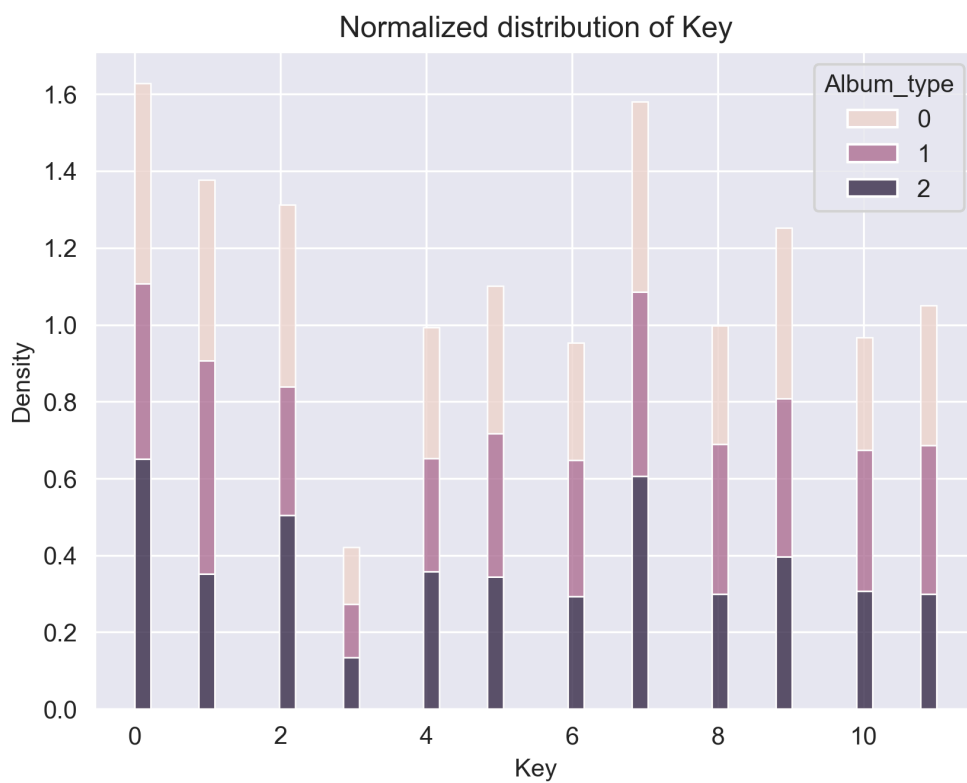
The popularity of a song key



Key mappings:

- **0** = C
- **1** = C \sharp /D \flat
- **2** = D
- **3** = D \sharp /E \flat
- **4** = E
- **5** = F
- **6** = F \sharp /G \flat
- **7** = G
- **8** = G \sharp /A \flat
- **9** = A
- **10** = A \sharp /B \flat
- **11** = B

Based on the histogram above, it can be seen that the least popular song key is the **D \sharp /E \flat** key and the most popular are **C**, **G** and **C \sharp /D \flat** , with **C** being the most frequently chosen one by a small margin.

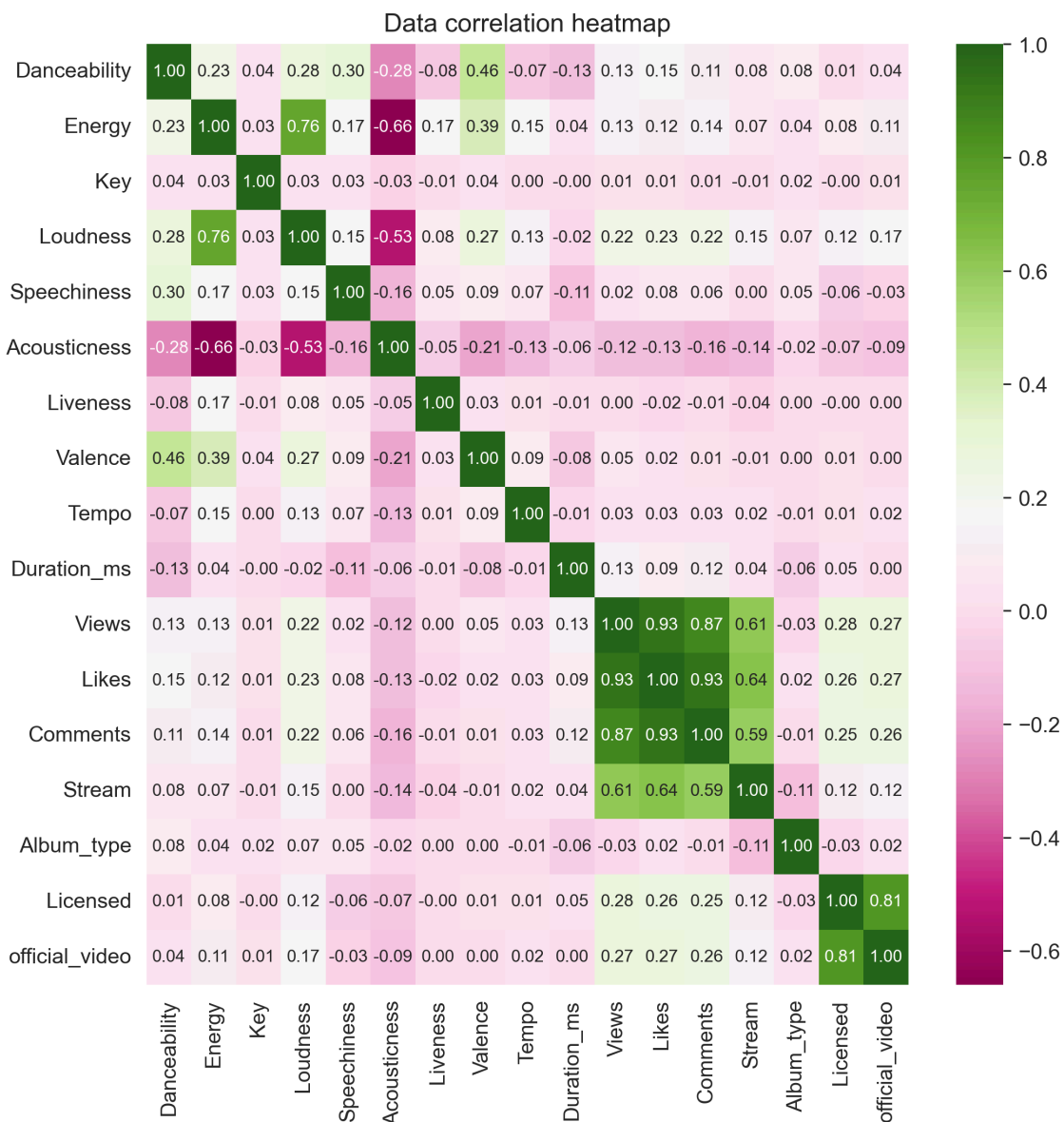


It is also worth noting that the independently normalized distribution with the album type hue shows that all keys have roughly equal popularity across different album types for each key.

Data correlation summary for the clipped dataset

To find which metrics are correlated, the best course of action is to prepare a correlation heatmap and focus on the metrics which have values on intersection point nearing one or minus one.

Correlation heatmap below:



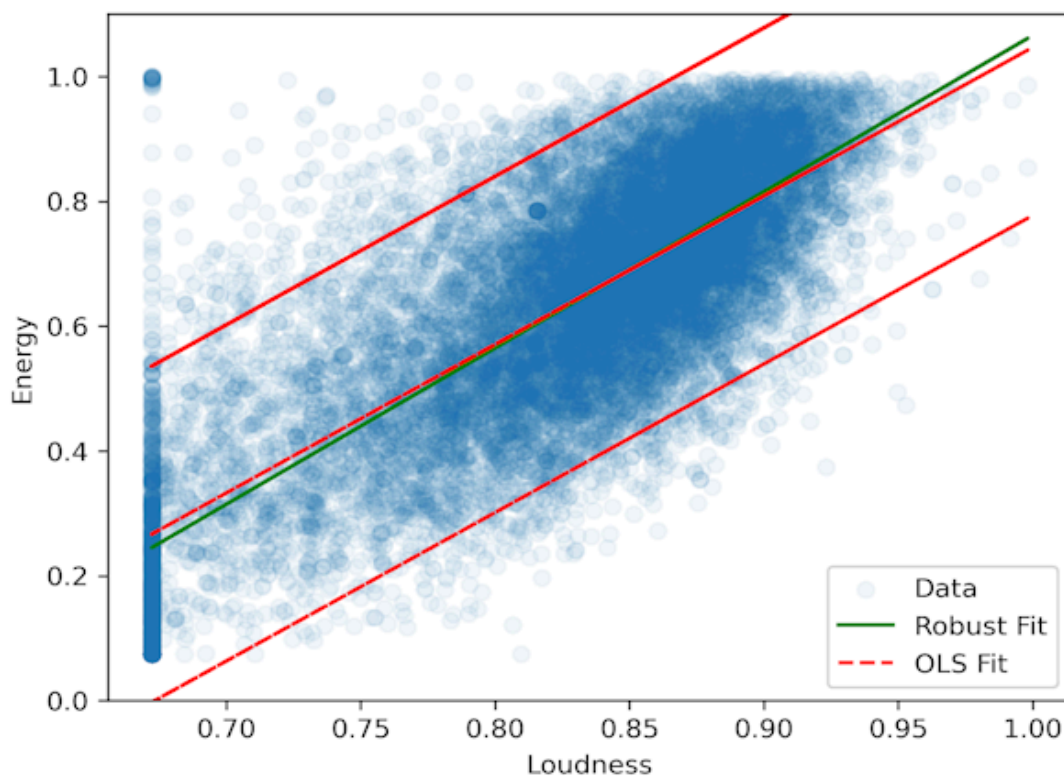
At first glance, five correlation groups can be identified:

- POSITIVE: Loudness and Energy Louder songs score higher in terms of song energy rating. This correlation will be looked into with more detail later on.
- POSITIVE: Licensing of the song and presence of an official video Vast majority of licensed songs have an official video uploaded to YouTube platform.
- POSITIVE: Count of views on YouTube and count of comments on YouTube and count of likes on YouTube With very strong positive correlation between the three metrics, it is a sign that they might be replaceable by fewer new metrics. Especially since they are all connected to songs performance on YouTube platform. This will also be investigated in more detail in the further parts of this summary.
- NEGATIVE: Acousticness and Energy Songs perceived to be more acoustic score lower in terms of song energy rating.
- NEGATIVE: Acousticness and Loudness Louder songs are less acoustic.

Detailed Loudness and Energy correlation visualization

To visualize and check if the correlation between the two metrics is, in fact, linear, two linear regression methods will be used.

- OLS regression — for standard visualization and prediction of lower and upper bounds
- RLM (Robust Linear Model) — more computationally intensive, but also included to show results with outliers down-weighted during regression



Both methods show comparable results that overlap with the densest area in terms of individual points on the plot. The vast majority of the points also lie within the OLS lower and upper bounds, which is a clear sign that the correlation between Loudness and Energy is linear.

Replacing YouTube views, comments and likes with a new metric

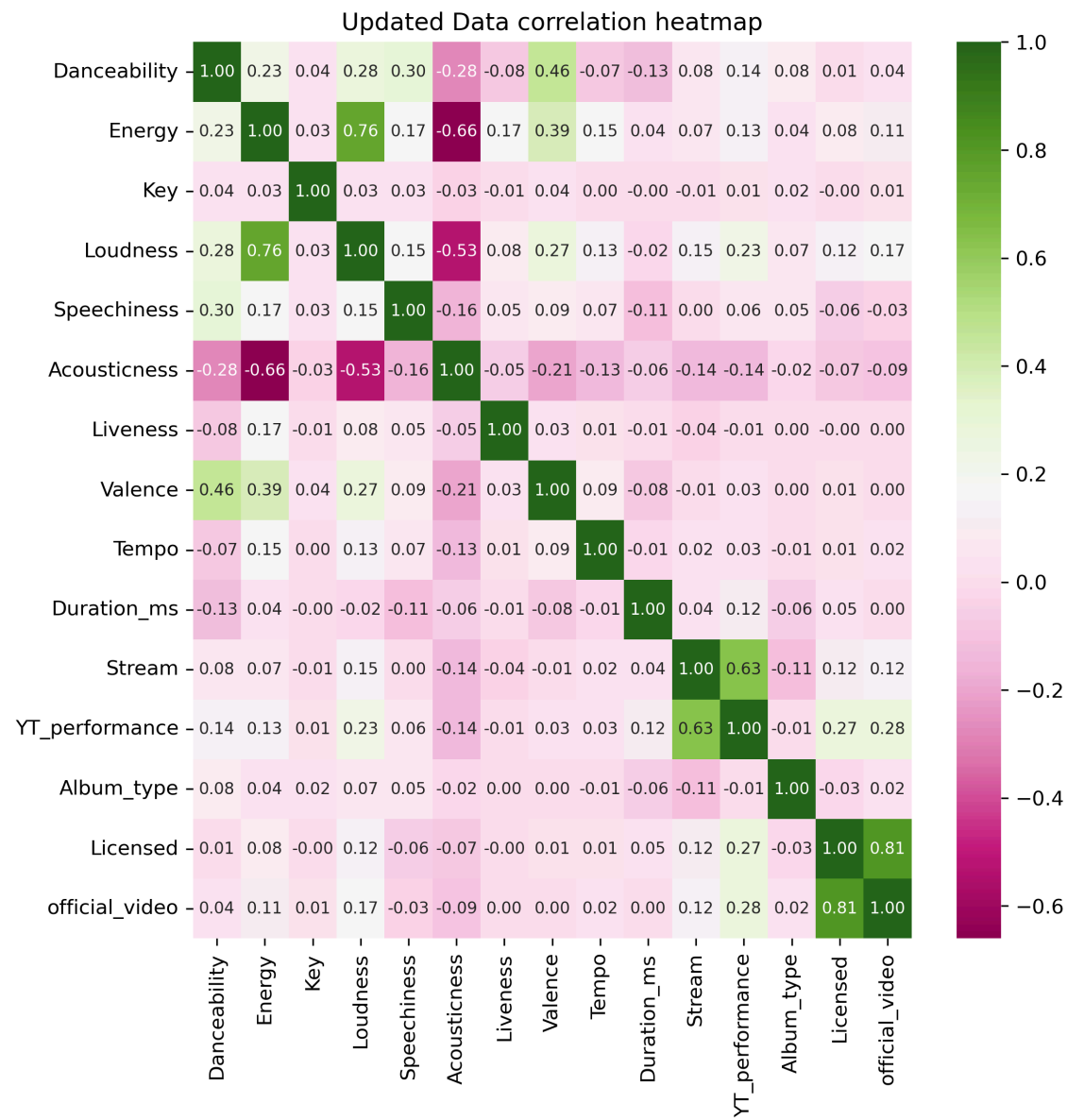
Assumption based on correlation heatmap: YouTube views, comments and likes share a strong linear collinearity and can be replaced

Chosen method for reducing the dimensions of correlated linear variables and creating a new metric is Principal Component Analysis (PCA). The First step is to choose the number of principal components. It will be determined based on eigenanalysis with two criteria:

- keeping only variables with eigenvalues above 1.0
- choosing the smallest number of variables while keeping the variance above 90%

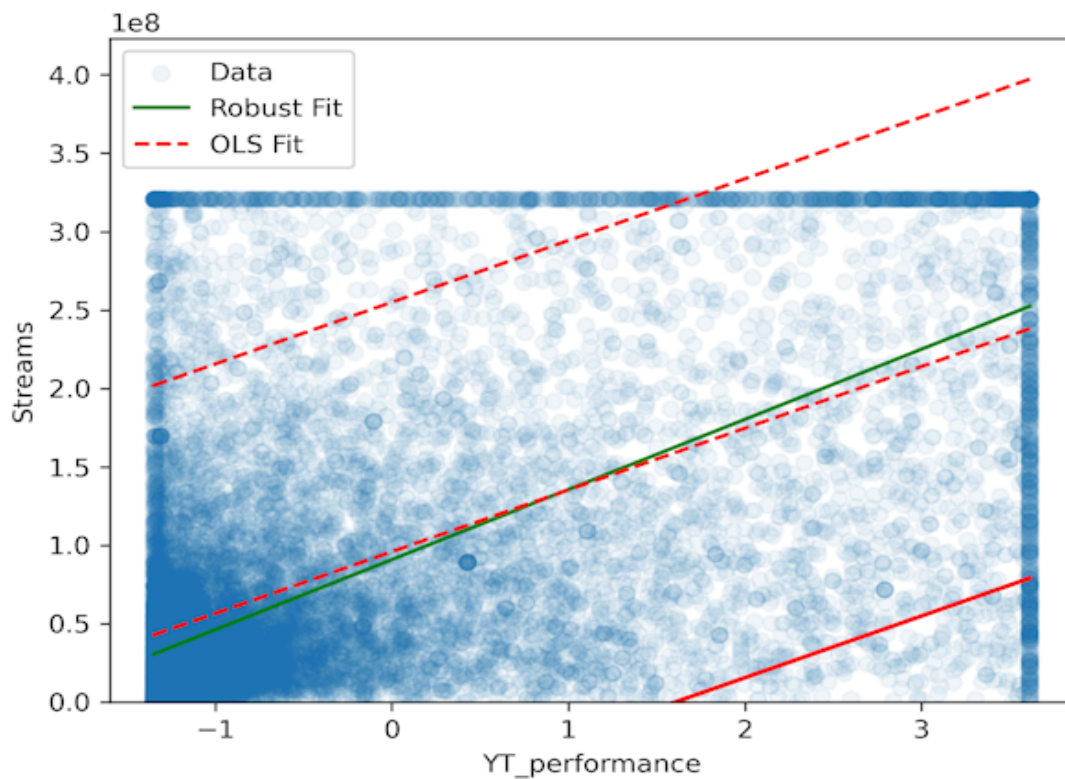
Detailed results of the eigenanalysis can be found in "eigenanalysis.csv" file. From the analysis, it can be seen that we need just the first principal component to satisfy the requirements. The new metric will be named "YT_performance" (songs performance on YouTube platform)

Updated correlation heatmap after PCA reduction:



Detailed Spotify and YouTube performance visualization

Using the same two OLS and RLM methods to draw a plot:



There are three main pieces of information that can be obtained from the plot:

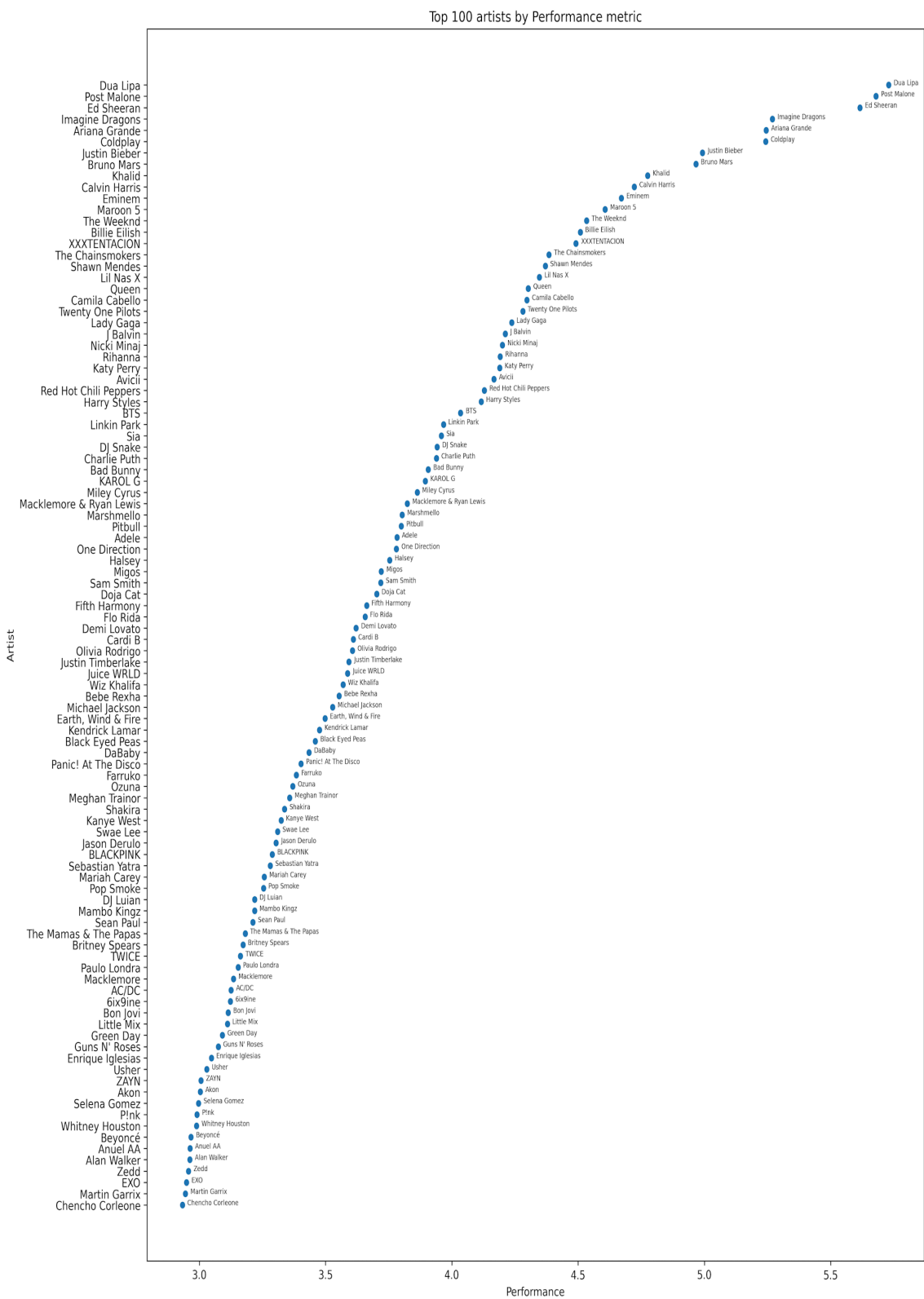
- There are artists who primarily share their works on Spotify, even to the point of ignoring YouTube altogether
- There are artists who do the complete opposite, preferring to work solely on YouTube
- In most cases, the popularity of a song is similar on both platforms, with less popular songs slightly more dominant on YouTube and the most popular songs having a larger presence on Spotify

Summary of top artists

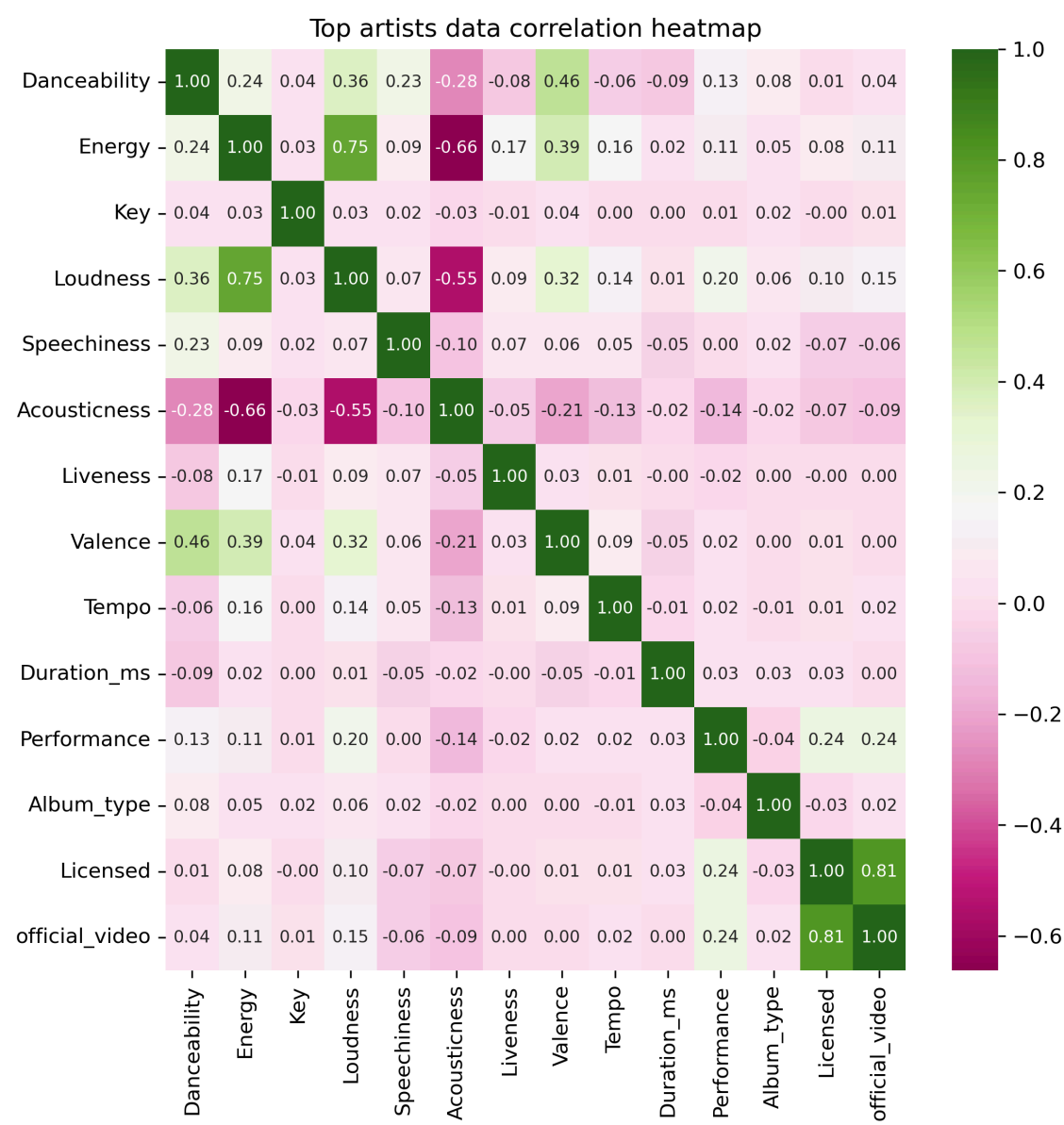
Creating a new "Performance" metric

Performance metric is based on the songs' performance on Spotify (count of streams) and YouTube ("YT_performance" metric) equal to $0.7 * \text{standardized_streams} + 0.3 * \text{standardized_YT_performance}$. The new feature takes performance on Spotify with more weight to reduce the impact of visual part of the song.

The plot below shows the summary of top 100 artists based on cumulative performance metric.



Additionally, it may be insightful to review correlation of metrics based on the small sample of the best performing hundred artists.



When compared to the previous heatmap from the entirety of the dataset, it can be seen that there are barely any changes in terms of correlation values. It signifies that the relationship between variables does not vary for the top artists and their less popular counterparts.