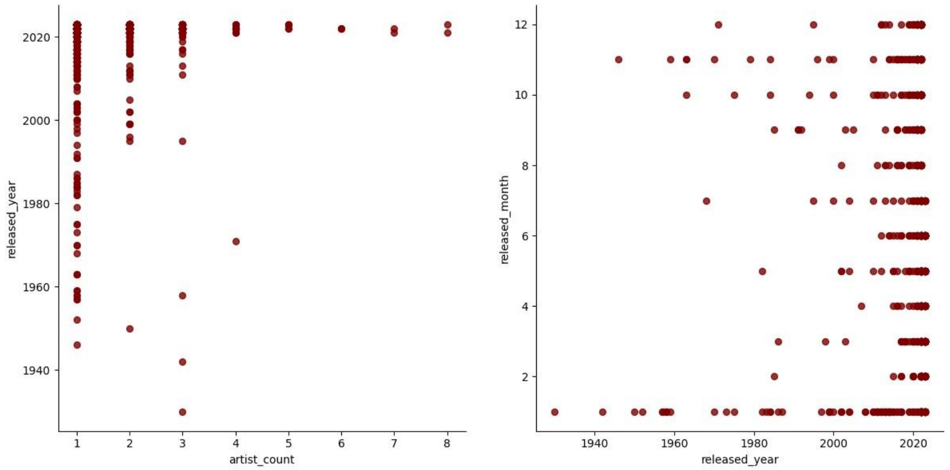
Exploring Music Trends: Machine Learning and Visualizations with Spotify Data

Harvard Undergraduate Data Analytics Group Fall 2023 Final Deliverable

Justin Liu

Introducing the Dataset



General Dataset Details:

- Most songs included are from 2020 and were released in January
- Dataframe shape: 953 x 21
- Several cases of "dirty data" (solved via imputing the missing/non-numeric values so as to not drop the data entirely

Artists vs # Songs Freq. Table

Artist Count	Number of Songs
1	587
2	254
3	85
4	15
5	5
6	3
7	2
8	2

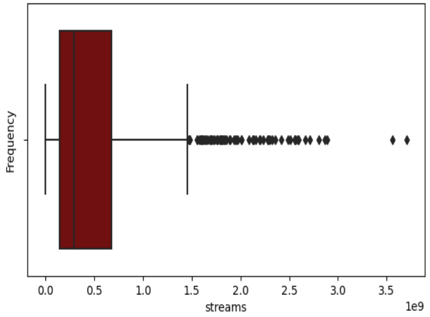
587 songs are attributed to a single artist, 254 songs are attributed to two artists, 85 songs involve three artists, and the remaining songs involve collaboration between four to eight artists.

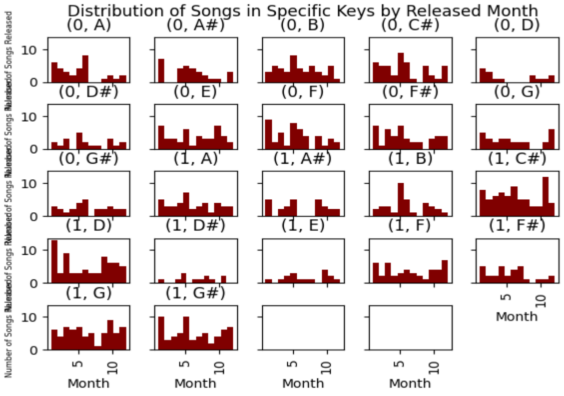
Variables of Data Type "Object" (non-integers):

- track_name
- artist(s)_name
- streams
- in_deezer_playlists
- in_shazam_playlists
- key
- mode

Exploratory Data Analysis and Pearson Coefficient Correlations

Checking the distribution and outliers for each column in the dataset, it was found that most of the data was positively skewed. This could have led to regression-based model difficulties in making accurate predictions, as it was forced to deal with rare cases and extreme values.

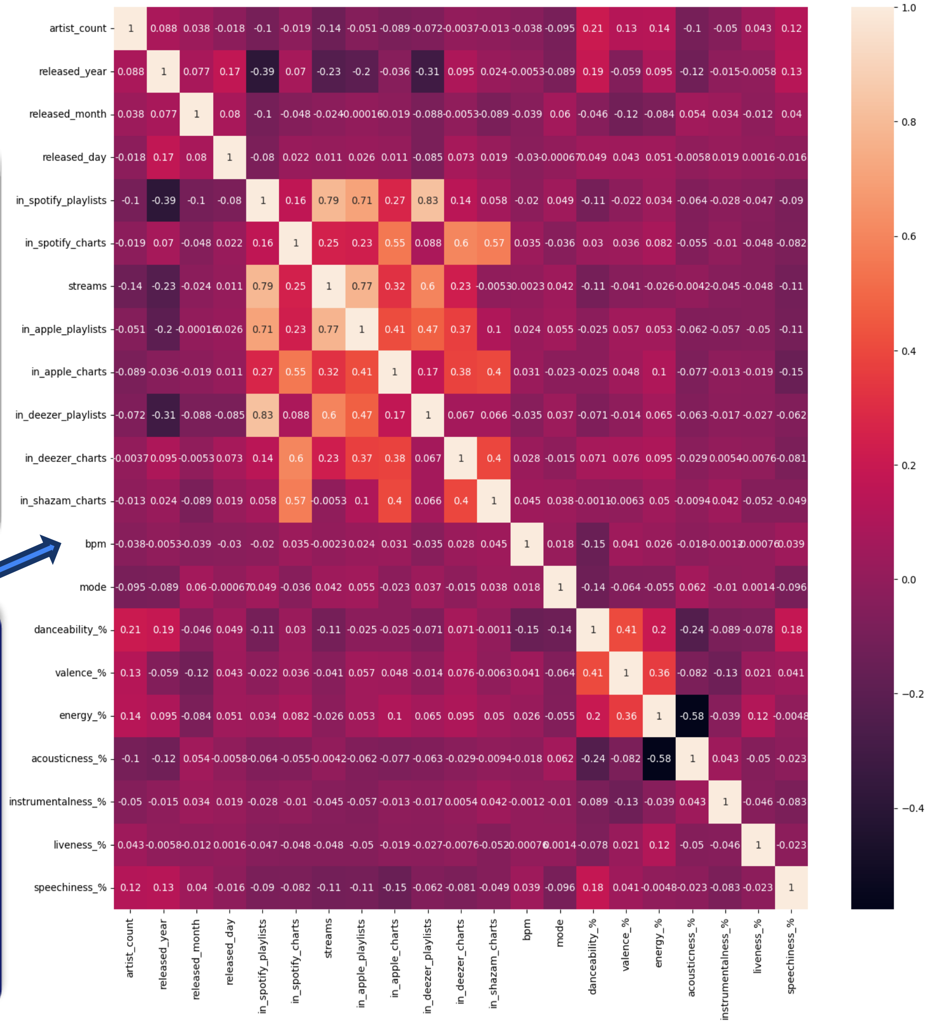




Seasonal patterns of song releases based on musical keys.

Pearson Coefficient Correlation Matrix

in_shazam_playlists show far lower correlation to either Spotify or Deezer, suggesting that these are 2 platforms that are more preferred by users.



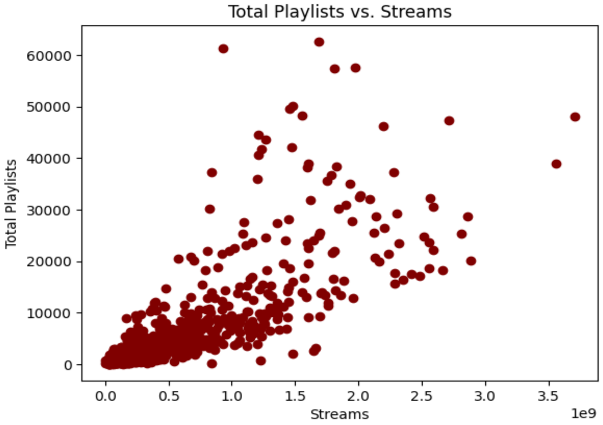
Initial Relationships: Artist Insights

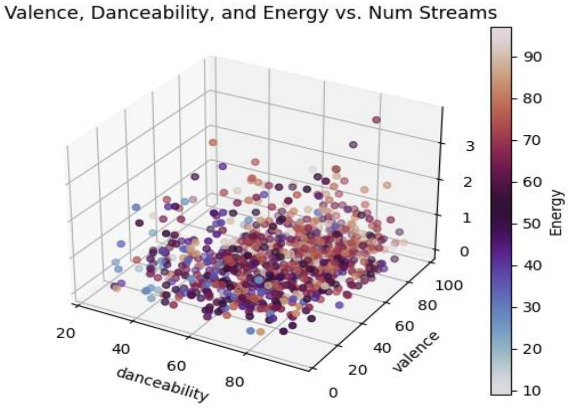
Number of songs by 5 random artists (including collabs)

{'Jung Kook': 2, 'Olivia Rodrigo': 7, 'Burna Boy': 2, 'Selena Gomez': 1, 'Myke Towers': 1}

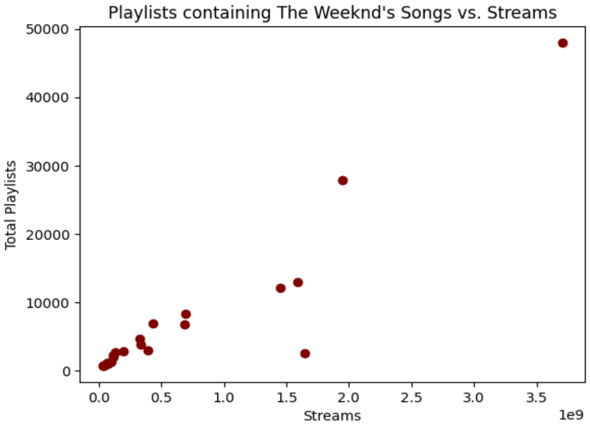
Top 5 artists identified by the number of songs in this dataset

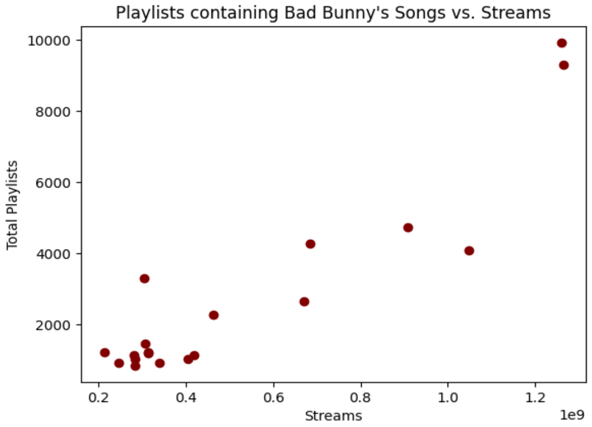
{'Bad Bunny': 40, 'Taylor Swift': 38, 'The Weeknd': 37, 'SZA': 23, 'Kendrick Lamar': 23}





- Comparing an artist such as The Weeknd to Bad Bunny, for example, we see that The Weeknd's correlation coefficient is much higher, and thus shows how he more consistently produces hit songs and possesses a more loyal fanbase.
- The plot illustrates a positive correlation between the song's number of streams and the number of playlists it appears in. This observation aligns with our intuition, as popular songs are more frequently included in playlists.





Is it possible to visualize a high-dimensional dataset in a human-interpretable way and reduce its complexity while capturing the most significant variation?

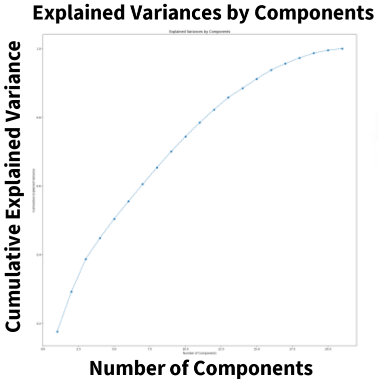
Principal Component Analysis (PCA)

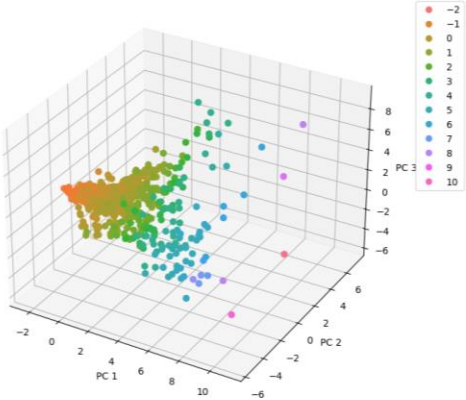
Following data **mean normalization**, compute the **covariance matrix** of the training data X:

$$\Sigma = \frac{1}{m}\sum_{i=1}^{m}(x^{(i)})(x^{(i)})^T$$

Then, the **eigenvectors** of the covariance matrix were found via **singular value decomposition**, where U is a matrix whose columns are the left singular vectors of X, and W is a matrix whose columns are the right singular vectors of X.

$$X = U\Sigma W^T$$





t-Distributed Stochastic Neighbor Embedding (t-SNE)

Equations on the left are initial **conditional probabilities**, final **joint probability** from **Gaussian Distribution** shown on the right. n= # of dimensions

$$p_{j|i} = \frac{exp(-\|x_i - x_j\|^2 / 2\sigma_i^2)}{\Sigma_{k\neq i} exp(-\|x_i - x_k\|^2 / 2\sigma_i^2)}$$

$$p_{i|j} = \frac{exp(-\|x_j - x_i\|^2 / 2\sigma_j^2)}{\Sigma_{k\neq j} exp(-\|x_j - x_k\|^2 / 2\sigma_j^2)}$$

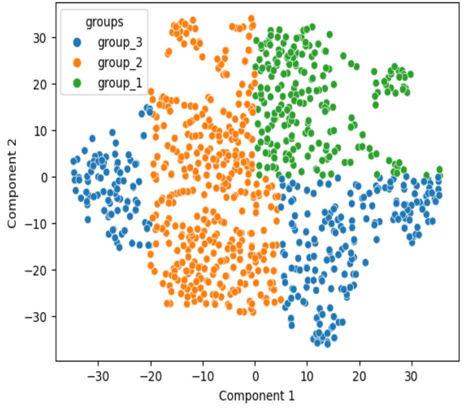
$$\longrightarrow \quad p_{ij} = \frac{p_{j|i} + p_{i|j}}{2n}$$

Second probability in **low-dimensional space**, given after **t-distribution**, yi and yj refer to points chosen to measure the distribution.

$$q_{ij} = \frac{(1 + \|y_i - y_j\|^2)^{-1}}{\sum_{k\neq l}(1 + \|y_k - y_l\|^2)^{-1}}$$

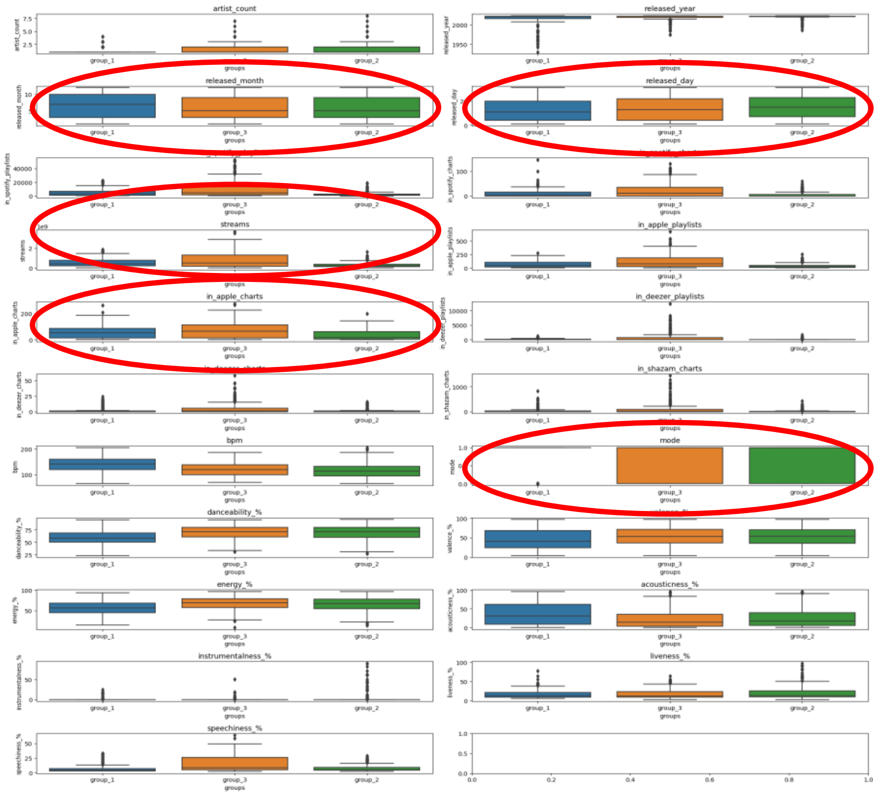
Final **cost function** measured through gradient descent, minimizes **Kullback-Leibler Divergence** between P and Q, the joint probability distributions of the high and low dimensions.

$$C = KL(P\|Q) = \sum_i \sum_j p_{ij} \log \frac{p_{ij}}{q_{ij}}$$



(cont.) How do we identify which attributes are the most variable across all samples using t-SNE clustering? Which attributes are the most important?

t-SNE Cluster Boxplots for all Variables



= high variation

Permutation Importance

Goal: Predict which statistics are most important in predicting the number of streams a song receives

Weight	Feature
0.0034 ± 0.0014	bpm
0.0031 ± 0.0033	liveness_%
0.0031 ± 0.0041	danceability_%
0.0022 ± 0.0022	valence_%
0.0020 ± 0.0029	streams
0.0020 ± 0.0022	in_apple_charts
0.0011 ± 0.0011	speechiness_%
0.0011 ± 0.0021	in_spotify_playlists
0.0011 ± 0.0021	in_spotify_charts
0.0011 ± 0.0021	in_deezer_playlists
0.0011 ± 0.0033	in_shazam_charts
0.0008 ± 0.0022	energy_%
0.0006 ± 0.0022	released_year
0.0006 ± 0.0014	instrumentalness_%
0.0003 ± 0.0011	in_apple_playlists
0 ± 0.0000	acousticness_%
0 ± 0.0000	released_month
0 ± 0.0000	released_day
0 ± 0.0000	mode
0 ± 0.0000	in_deezer_charts

… 1 more …

Highest-impact features

Procedure

1) Train a basic model to predict # streams based on the song's statistics.

2) Shuffle the values in a single column and make predictions using the resultant dataset. Compute the loss function to measure the importance of the variable just shuffled.

3) Return the data to the original order. Repeat step 2 with the next column in the dataset until all variable importances have been calculated.