Exploring Musical Emotion

Introduction

Music is deeply emotional, and the emotions it conveys are often quantified using structured data. Spotify's valence score attempts to do just that. It measures the musical "positivity" of a track on a scale from 0 (sad, dark, or angry) to 1 (happy, euphoric, or cheerful). This exploratory analysis will try and figure out if this valuence score is a subjective quality or a predicted and objective measure. This analysis will focus on spotify song data on top hits. Ultimately, this investigation aims to build an accurate valence predictor, and to reflect on the broader question: To what extent can machines understand the emotional language of music?

Overview of the Dataset

This dataset consists of 2,017 tracks, each described by a range of audio features generated through Spotify's music analysis algorithms. These features quantify aspects of rhythm, loudness, instrumental presence, and more. Key attributes include:

- Danceability: how suitable a track is for dancing, based on tempo, rhythm stability, and beat strength
- Energy: the intensity and activity level of a track
- Loudness: the average decibel level across the entire track
- Acousticness: the likelihood a track is acoustic
- Speechiness: the presence of spoken words
- Mode: whether the track is in a major (happy) or minor (sad) key
- Valence: a numeric score between 0 and 1 indicating the track's emotional positivity

Each track is also assigned a key (C, D, E, etc.), a time signature, song title, and artist which are treated as categorical features.

Predictive Task

The primary objective of this analysis is twofold.
First is to explore the relationships between audio features and musical mood. Second is to develop regression models that can predict the emotional positivity of a song, using Spotify's track-level audio data.

What is the point of this project? - To extract my own observations from manipulation of this dataset and trying to answer these questions:

- How is musical valence distributed across the dataset?
    - Are happier songs overrepresented on streaming platforms?
- Which audio features most strongly correlate with valence?
    - Are rhythmic features more important than tonal ones?
- How does valence differ between major and minor modes?
    - Is mode a reliable indicator of emotional tone?
- Are loud songs perceived as more positive — or does positivity peak at moderate volumes?
- What feature interactions reveal distinct emotional identities in music?
- Can I accurately predict valence using just objective audio features?
    - How close can I get to Spotify's own valence scores?
- Which model architecture best captures emotional signals in the data?
    - Do non-linear models like Random Forest outperform linear regressors?
- How does tuning model hyperparameters affect accuracy?
- What is the lowest achievable MSE when predicting valence?
    - And how does that compare to a naive average-predicting baseline?
- What limitations prevent perfect prediction of emotional tone?

Could incorporating lyrics, genre, or listener feedback improve model performance?

The project is organized as follows:

I. Data Cleaning
II. Exploratory Data Analysis

- Feature Distributions
- Valence Relationships by Mode and Key

III. Feature Selection and Engineering
IV. Modeling

- Linear Regression
- K-Nearest Neighbors
- Decision Tree
- Random Forest
- Gradient Boosting

V. Summary and Conclusion

I. Data Cleaning

The Spotify dataset used in this project contains over 2,000 tracks, each described by detailed audio features extracted by Spotify's music analysis algorithms.

Initial Data Assessment
Upon inspection, the dataset displayed several strengths:

- No missing values across any key feature columns
- Consistent data types (floats for continuous features, integers for categorical ones)
- Well-structured audio features that align with Spotify's internal metric definitions

This is seen as the dataset chosen was released officially by Spotify. Hence, it is a pre-engineered dataset that doesn't contain inconsistencies you would expect from sensor based data, or manually inputted data that could have user error inputs. It also does not contain any missing values as every song has a value for each feature. It is effectively impossible for a song not to have a datapoint for the features mentioned above.

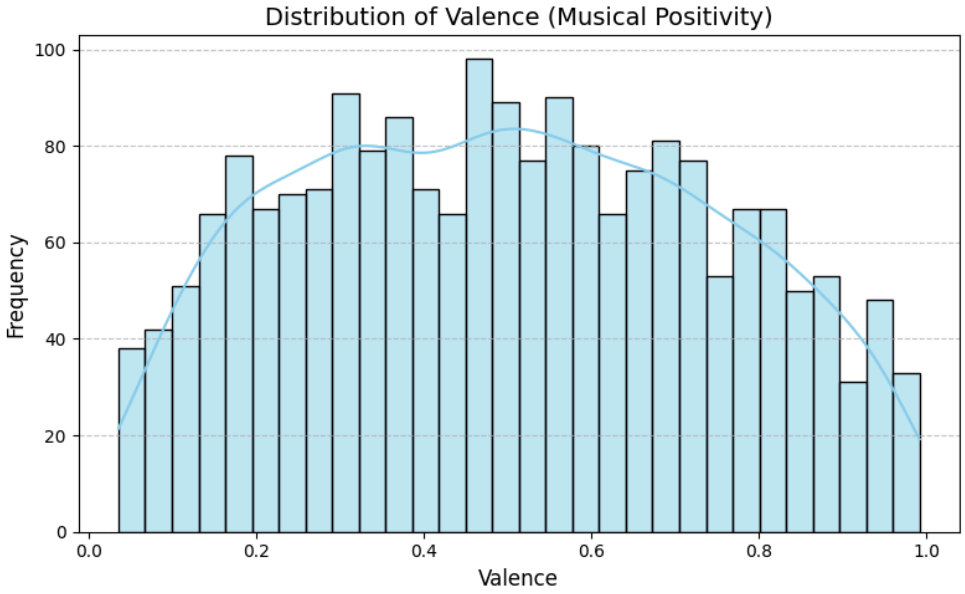
Data Cleaning Steps

- Column Reduction
    - Dropped extraneous columns such as *Unnamed: 0, song_title*, and *artist* that were irrelevant for analysis or modeling.
- Duplicate Handling
    - Identified and removed 6 duplicate records using .drop_duplicates() to ensure each song is uniquely represented.
- Categorical Mapping
    - Converted numerical representations of the *key* column (0–11) into musical note labels (e.g., 'C', 'D#', 'F') to improve interpretability during visualizations and model evaluation.

Data Storage and Reproducibility
All cleaning steps were performed within a reproducible Python pipeline using *pandas*. The cleaned dataset was maintained within a single version-controlled notebook to preserve transparency and enable consistent transformation across both exploratory and modeling notebooks.

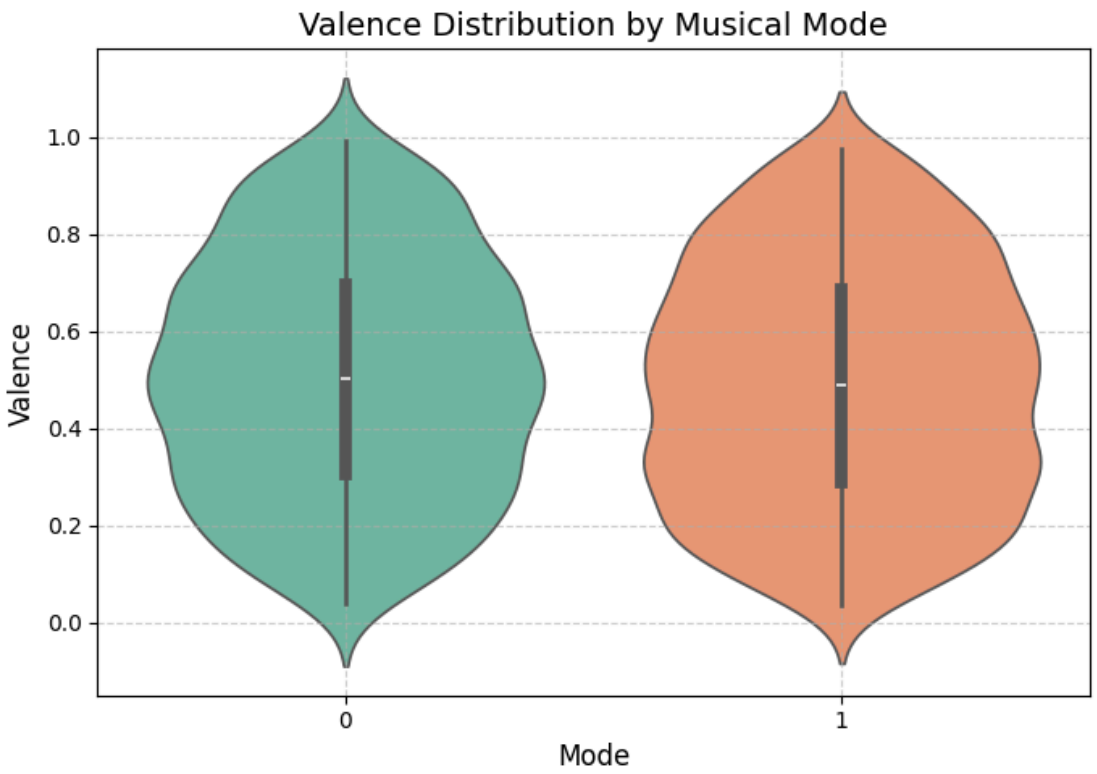
II. Exploratory Data Analysis

I began the analysis by visualizing the distribution of valence across the dataset.



The result was a relatively even spread across the 0–1 range, with a very slight concentration toward the center. This suggests that most songs avoid emotional extremes. They are neither excessively melancholic nor euphoric. I think that this reflects the broader listener preferences or platform curation. It also indicated that valence is a usable target variable with enough variance to be meaningfully predicted.

I then explored how the valence varies by the musical mode.

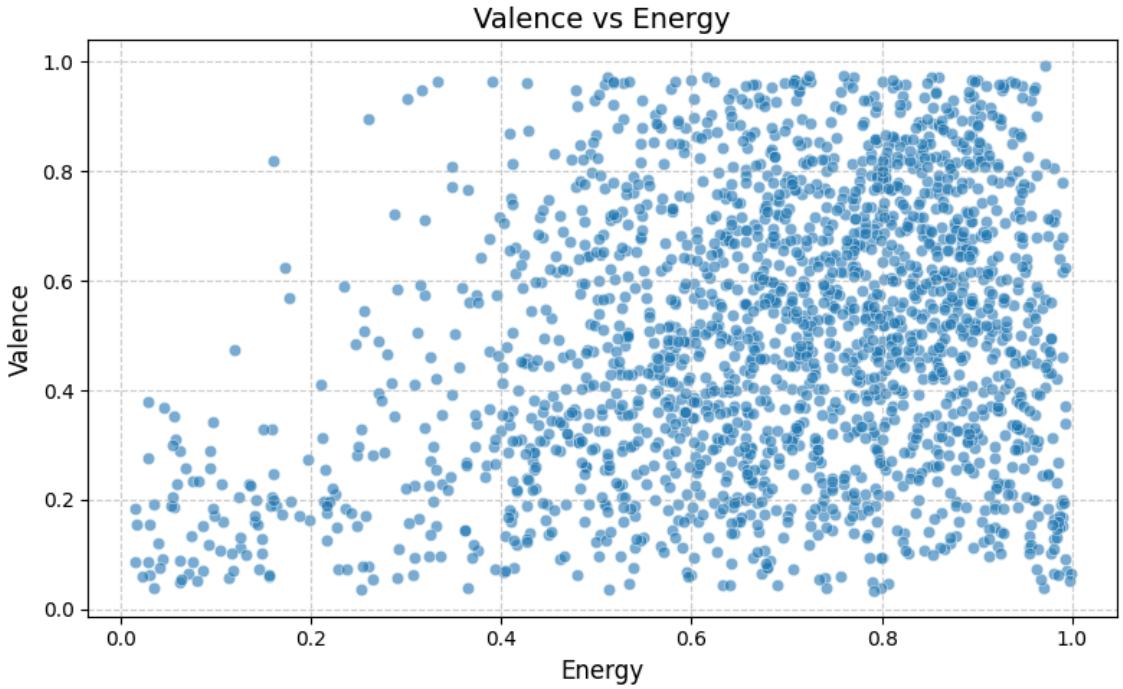


Songs in major keys had, on average, higher valence than those in minor. This is somewhat notable but the effect size was modest. A violin plot reveals large internal variance within each group, indicating that while mode can align with general emotional trends, it isn't a standalone determinant of mood.

A similar pattern was seen in valence by key, where keys like A and B skewed slightly higher, but again the differences weren't decisive. These observations suggested that both mode and key should be retained as categorical variables for modeling, but are unlikely to be dominant predictors.

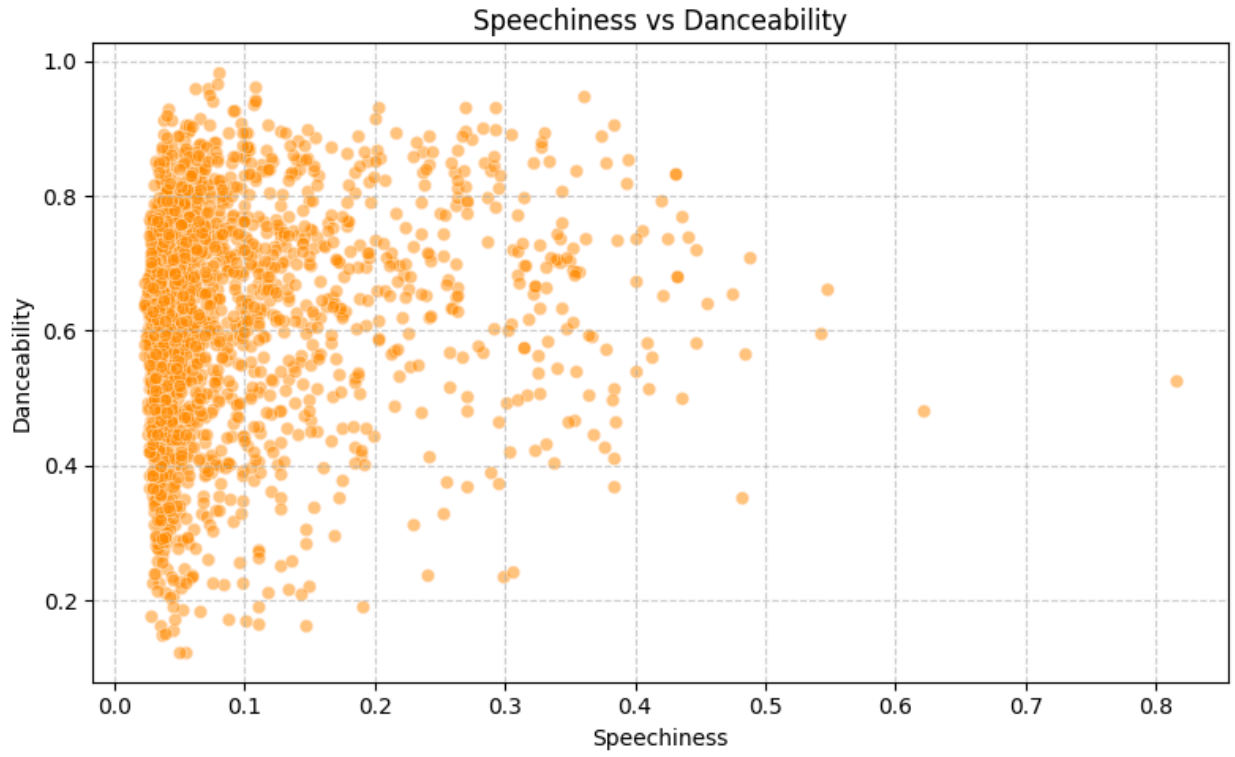
Next, I looked at feature-to-target correlations. A heatmap showed that energy had the strongest positive correlation with valence at (~0.38). This suggests that tracks with acoustic instrumentation are more likely to be perceived as somber or introspective. Features like speechiness and instrumentalness showed little linear correlation, but were retained due to genre-specific variability observed in scatterplots.

Furthermore, I looked at some of the bivariate scatterplots. I used these to help find if there was any visual correlation between any variables. This would help me narrow down which variables I could use for the modelling section of my project. A scatterplot between Energy and Valence confirmed the heatmap findings, showing a clear upward trend. The high-energy songs clustered around higher valence values.



In contrast, a plot of instrumentalness vs. loudness demonstrated a large range in loudness for instrumental tracks. This showed me that the emotional tone in these songs relies more on volume dynamics than lyrical content. This was slightly unrelated to my overall goal of this dataset exploration, but was an interesting insight.

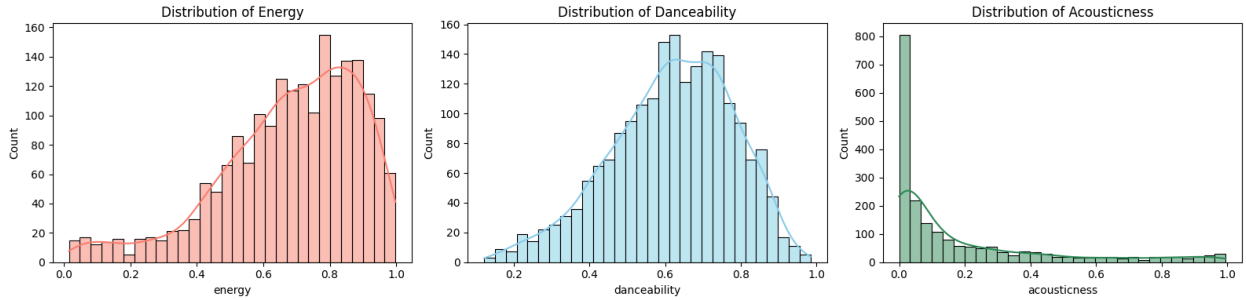
Another key insight came from danceability vs. speechiness, where genre-based clustering became apparent.



Tracks with high speechiness and moderate danceability likely represented rap or hip-hop, and showed high variance in valence. This showcases the need for models that can capture non-linear relationships, particularly when genre proxies like speechiness can introduce emotional ambiguity about songs.

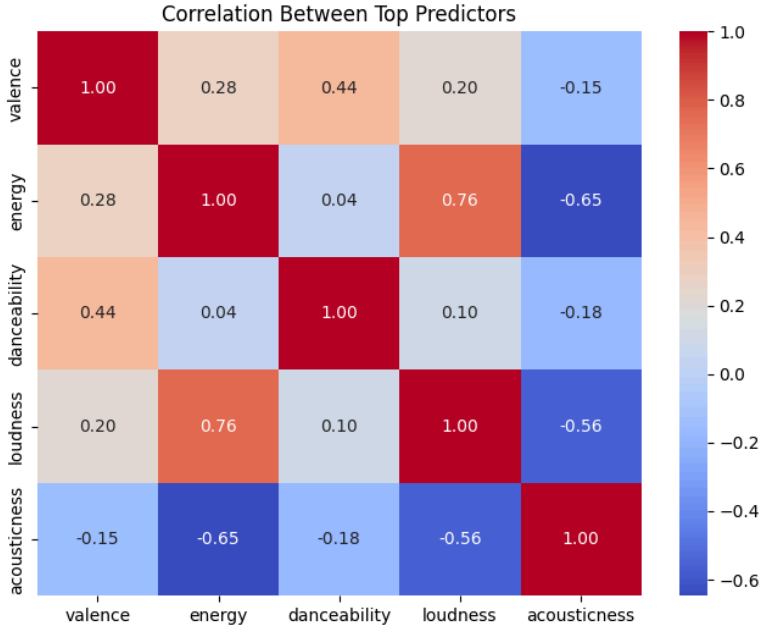
Looking at loudness by musical key, I observed that certain keys (C#, D, and A) tended to support louder tracks, while others (F, G) leaned quieter. While not conclusive on its own, this pattern informed my choice to preserve key as a modeling feature.

Tempo, surprisingly, had almost no meaningful correlation with valence. The scatterplot confirmed the flat relationship, which cut across tempo ranges. This finding challenged the assumption that faster songs are inherently happier, and led us to deprioritize tempo in feature selection for simpler models, though it was kept in ensemble approaches in case of interaction effects.



Examining the distributions of my top predictor variables provided important context.

Energy skews slightly right, with most tracks falling between 0.6 and 0.9. This showed a strong bias in modern streaming content toward high-intensity music. This aligns with the earlier finding that energy has the strongest positive correlation with valence. Danceability is more evenly distributed, with a slight peak around 0.7, suggesting that most tracks are rhythmically stable and listener-friendly. In contrast, acousticness is highly skewed left, meaning the majority of tracks are not acoustic. This supports the earlier correlation finding: acoustic songs tend to have lower valence, but they're also underrepresented in the dataset, which may influence model training. These insights help explain both the predictive power and the practical limitations of each feature.



Beyond looking at how each feature relates to valence, I also examined correlations between the predictors themselves. Notably, energy and loudness showed a very high correlation (~0.75), suggesting that they capture overlapping dynamics in the dataset. This could raise concerns about multicollinearity in linear models, though it is less

problematic in tree-based approaches. Danceability showed moderate positive correlations with both energy and loudness, indicating that more energetic songs also tend to be more danceable. Acousticness, interestingly, was negatively correlated with all three: energy, danceability, and loudness. This confirms that acoustic songs are not only quieter and calmer, but also rhythmically unique. These findings provide a more holistic understanding of how audio traits interact.

By the end of the EDA, I identified energy, danceability, loudness, and acousticness as the most promising predictors of valence. Mode, key, and speechiness were retained for categorical diversity and potential genre-related variance. Tempo and instrumentalness were deprioritized but not discarded outright.

III. Feature Selection and Engineering

1. Tonality and Key-Based Encoding

Musical key and mode are categorical features that capture important tonal elements of music. While my analysis showed that key alone does not determine valence, some keys (like A and B) showed higher average loudness and valence. We encoded both key (as a 12-note chromatic scale from C to B) and mode (major/minor) using one-hot encoding to avoid introducing ordinal bias.

This transformation allowed me to keep small stylistic signals that different artists can have, while enabling compatibility with linear and tree-based models. It also preserved genre distinctions indirectly encoded in harmonic structure (e.g., minor keys common in hip-hop vs. major in pop).

2. Rhythmic Signal Scaling

Features like energy, loudness, tempo, and danceability reflect different rhythmic and production qualities of a track. The EDA showed that energy had the highest correlation with valence (+0.55), followed by loudness and danceability. Tempo, however, was weakly correlated and included primarily to allow interaction terms to capture compound effects in ensemble models.

To standardize these features across varying scales, we applied min-max normalization for visual analysis and z-score standardization during modeling, particularly for KNN and neural net pipelines that are sensitive to scale.

3. Feature Interaction Engineering

Currently, my model would treat predictors independently. However, emotional tone in music often emerges from the interaction between features — for example, high energy combined with low acousticness may show a vibrant electronic track, while high speechiness and loudness may suggest rap. To capture these compound effects, I tried to engineer interaction terms between top predictors such as energy × danceability, or loudness × instrumentalness.

These nonlinear combinations could be particularly valuable in boosting performance for linear models and may help expose genre structures that could not be explored from individual features.

IV. Modelling

Having established valence as a rich emotional target and identified energy, loudness, danceability, and acousticness as its strongest predictors, I trained a series of regression models. Our goal was not just to minimize prediction error, but to explore whether models can reliably translate objective audio traits into subjective emotional interpretations.

All models were evaluated using Mean Squared Error (MSE), with a naive mean-prediction baseline serving as our benchmark. Importantly, these modeling efforts also allowed us to revisit and test several of our original exploratory questions, including how emotional tone is structured, whether linear assumptions hold, and which modeling frameworks are most effective. Valence, our target variable, is continuous and bounded between 0 and 1, so even small changes in MSE represent meaningful increases/decreases in accuracy.

Baseline
Our baseline model predicted the mean valence score for all test samples. This model, while intentionally simplistic, provides a crucial performance floor:

MSE: 0.0627
RMSE: 0.2504

The relatively high RMSE confirms that valence is not a uniform outcome and varies meaningfully across tracks. This alone answers one of our early questions: yes, valence has enough variance to be predicted.

Decision Tree Regressor
Decision trees capture nonlinear thresholds in the data, ideal for handling categorical splits like mode or key. Initial results were poor due to overfitting. After tuning max_depth and pruning the tree, performance improved:

MSE: 0.0503

What makes the decision tree interesting is interpretability. I visualized the structure and extracted feature importance, which reaffirmed that energy, danceability, and acousticness consistently appeared near the top of the splits. This partially answers my earlier question on which features most strongly drive emotional variation, confirming what EDA had hinted.

Still, the model underperformed compared to more robust alternatives, and residual errors remained large for emotionally extreme tracks (very low or high valence).

K-Nearest Neighbors (KNN)
KNN predicts valence by averaging the target values of the k closest examples in feature space. This method relies heavily on local similarity, assuming that songs with similar features have similar emotional tones.

MSE: 0.0751
KNN performed worse than expected. While conceptually appealing (a sad, acoustic song should be close to other sad, acoustic songs), the high dimensionality and noisy variance in valence made the neighborhoods less reliable. This failure suggests that musical emotion is not locally clustered in a consistent way. This answers another of my questions: proximity in feature space does not guarantee emotional similarity.

Multiple Linear Regression
The Linear regression was surprisingly very strong.

MSE: 0.0438

The model performed well, especially in the mid-valence range. This reinforces my hypothesis that valence has a substantial linear component, particularly through features like energy and loudness. However, residual plots showed poor performance at the extremes (0–0.2 and 0.8–1), where emotions are most intense or ambiguous.

This supports the insight that extreme emotional tones may depend on compound or nonlinear interactions (high energy with low acousticness and major key) that linear models can't capture directly.

Random Forest Regressor
Random forests use an ensemble of decision trees to reduce overfitting and generalize better. This model struck a strong balance between interpretability and flexibility.

MSE: 0.0417

Random forests offered significant gains over linear and KNN models. They could capture interactions without needing to explicitly engineer interaction terms, and handled outliers better. Importantly, they ranked acousticness and speechiness higher in importance than linear models did. This suggests they're meaningful in genre-specific moods, even if not the predictors are not individually linear.

This partially answers my question about nonlinear relationships: yes, tree-based models offer a meaningful advantage in capturing emotional interactions.

XGBoost (Tuned)
After grid search tuning (max depth, learning rate, n estimators), this is the XGBoost output:

MSE: 0.0411

XGBoost builds trees sequentially, correcting errors from prior models, making it ideal for complex regression problems. Its marginal improvement over Random Forest suggests my models are approaching the performance ceiling of the feature set. XGBoost's residuals were narrower and better distributed, particularly in the upper-valence range.

The model's strength in generalization confirms that the features do encode emotional signals well enough for strong prediction. However, it also shows that some variance in valence remains unexplained, possibly due to external factors (lyrics, cultural context, etc.).

Neural Network
Neural nets are designed to find deep patterns through weighted connections across hidden layers. We tested a two-layer neural network with 64 nodes per layer.

MSE: 0.0471

While powerful, the neural network slightly underperformed due to overfitting and limited data volume. With further tuning (dropout, activation functions, deeper layers), this model could improve. I tried using 3 layers and 4 layers to see if it would improve the performance, however, nothing changed. Its current underperformance could be an earlier insight: feature quality matters more than model complexity when the signal-to-noise ratio is low.

V. Conclusion

Emotional Audio Patterns: Key Observations

My analysis confirms that music valence, an emotional measure of positivity, is meaningfully correlated with measurable audio features. Consistent with the exploratory findings, songs with higher energy, danceability, and loudness tend to have higher valence, while those with elevated acousticness and instrumentalness often score lower. This highlights an important insight: even abstract emotional qualities like musical "happiness" can be modeled using objective, quantifiable data.

Predictive Modeling Highlights

Model Performance

My model shows a clear hierarchy of performance. The baseline model, which predicted the mean valence for all songs, yielded the highest MSE (0.0627), validating the non-triviality of the task. Linear models performed surprisingly well, suggesting a strong global trend in how energy and loudness affect positivity. However, more flexible models like Random Forest and XGBoost achieved lower MSEs, indicating their superiority in handling edge cases and nonlinear relationships. Neural networks and KNN, though conceptually promising, underperformed relative to ensemble methods.

Feature Interpretability

Across all models, energy, danceability, loudness, and acousticness repeatedly emerged as top predictors. The importance of these features aligns with how listeners intuitively experience emotional tone. This confirms that upbeat, loud, energetic tracks tend to have a positive effect. These findings address my earlier exploratory questions by affirming that:

1. Valence is not purely subjective, it can be predicted with some accuracy

2. Emotional tone is not evenly distributed, it clusters around structural traits like rhythm, instrumentation, and key
3. Some features (tempo, mode) are only weakly predictive alone but gain relevance when combined with others

By using models like this project has, artists can effectively find a ‘formula’ to making hits, as people will tend to listen to songs that make them feel happier. This is shown by the connection between emotion and audio features. By grounding subjective experiences like valence in measurable patterns, we offer a framework that can be applied in recommendation systems.