Feature Engineering Report – Avenged Sevenfold Songs Dataset

1. Dataset Chosen

The dataset used for this assignment is the **Avenged Sevenfold Songs Dataset** (Kaggle).
It contains song-level information including:

1 Album
2 Song name
3 Popularity score
4 Release date
5 Track number
6 Energy
7 Tempo
8 Time signature
This dataset is used to demonstrate feature engineering techniques on a real-world music metadata set.

---

2. Feature Engineering Steps Performed

A. Data Exploration

Inspected data types using `df.info()`
Viewed first rows using `df.head()`
Identified categorical features: *album*, *name*
Identified numeric features: *popularity*, *track_number*, *energy*, *tempo*, *time_signature*, *year*

B. Handling Missing Values

Numeric columns → **Median imputation**
Categorical columns → **Most frequent imputation**
Confirmed all missing values were removed

Reason:
Median handles outliers better than mean, and most-frequent maintains common categories.

C. Feature Encoding

Converted *album* and *name* into numeric format using **One-Hot Encoding**
Dropped the first category to avoid dummy variable trap

Reason:
Machine learning models cannot process text labels.

D. Feature Scaling

Applied **StandardScaler** to all numeric features:

Scales mean → 0
Scales variance → 1

Reason:
Helps models like PCA and KBest treat all features equally

E. Feature Extraction (PCA)

Applied **PCA with 5 components**
Reduced dimensionality while preserving maximum variance
Impact:
Dataset becomes compact, easier to model, and faster to compute.

F. Feature Selection

Used **SelectKBest (ANOVA F-test)** to select 5 best features related to our target.
Removed low variance columns using VarianceThreshold.
Purpose:
Improves performance & removes irrelevant features.

G. Target Variable Creation

Since the dataset had no target, we created a classification target:

Era Classification

0 = Old A7X Era (≤ 2013)
1 = Modern Era (> 2013)

This allows us to perform feature engineering in a supervised learning context.

3. Observations After Transformation

One-hot encoding expanded dataset (more columns)
PCA compressed dataset from many columns down to 5 high-value components
SelectKBest highlighted features like **year, tempo, energy, popularity** as influential
Scaling improved comparability between tempo/energy/popularity
Dataset became cleaner, more consistent, and ML-ready

4. Ethical Considerations

This dataset does not include sensitive personal information (gender, religion, ethnicity, marital status).
However, ethical concerns still apply:

Bias Issues

Popularity values may bias the model toward newer or more streamed songs. Song titles might introduce accidental bias if used for predictions.

Mitigation

Avoid using non-musical attributes (song name) when predicting quality.
Normalize popularity scores.
Ensure fairness by analyzing feature influence.

Conclusion

The dataset was successfully transformed using full feature-engineering workflow, including handling missing data, encoding, scaling, PCA, feature selection, and target creation.
The resulting dataset is now optimized for machine learning classification tasks and reflects strong preprocessing and data-cleaning practices.