# Spotify Song Clustering for Personalized Recommendations

## Project Overview

This project aims to enhance user experience by clustering similar songs to improve Spotify’s recommendation system. By grouping songs into cohorts based on relevant features, the system can provide users with personalized song suggestions. The project involves data analysis, feature selection, dimensionality reduction, and clustering techniques to achieve this goal.

## Problem Scenario

Spotify, a Swedish audio streaming and media service provider, has over 456 million active monthly users, including 195 million paying subscribers as of September 2022. To retain engagement, Spotify seeks to implement a song clustering mechanism based on different attributes of the songs. The clustering will help create curated song groups that improve recommendations and user experience.  
  
The objective is to utilize exploratory data analysis (EDA), principal component analysis (PCA), and clustering techniques to form meaningful song cohorts.

## Steps Undertaken

### 1. Data Cleaning and Preparation

- Imported the dataset and verified the integrity of song attributes.  
- Addressed missing values by imputing data based on similar song characteristics.  
- Standardized numerical features to ensure consistency across all attributes.

### 2. Exploratory Data Analysis (EDA)

- Generated visualizations such as histograms and box plots to analyze data distributions.  
- Identified outliers and performed necessary treatments to refine the dataset.  
- Used heatmaps to examine correlations between various song attributes.

### 3. Principal Component Analysis (PCA)

- Applied PCA to reduce dimensionality while preserving critical information.  
- Determined the optimal number of principal components to retain for clustering.

### 4. Cluster Analysis

- Implemented clustering techniques such as K-Means and DBSCAN.  
- Evaluated clustering performance using metrics like the Silhouette Score.  
- Identified key patterns in the song clusters to refine recommendations.

### 5. Data Visualization and Insights

- Analyzed song clusters to understand similarities and differences.  
- Visualized the impact of key song attributes on clustering outcomes.  
- Evaluated the efficiency of clustering techniques using data-driven insights.

## Key Takeaways

This project showcases the power of data-driven insights in enhancing Spotify’s recommendation system. By leveraging clustering techniques, Spotify can provide more personalized recommendations, ensuring an improved user experience. The combination of PCA and performance evaluation metrics strengthens the reliability of the clustering process.

## Technologies Used

- Python: Core language for analysis and modeling.  
- Pandas: Data cleaning, transformation, and management.  
- Matplotlib & Seaborn: Data visualization.  
- Scikit-learn: PCA, clustering techniques, and performance evaluation.

**Author**: Ravi Ranjan Lal

**Email**: lalraviranjan@gmail.com