Assignment #3: Music Data Analysis with Apache Spark

Objective: Use Apache Spark to perform exploratory data analysis on a dataset of music tracks. Analyze the attributes that contribute to track popularity and explore trends within music genres.

Data Exploration with Apache Spark (10%)

1. **Data Loading and Schema Understanding (2%)**
    - Load the dataset into a Spark DataFrame.
    - Print the schema and verify the data types of each column.
2. **Data Aggregation (3%)**
    - Calculate the average danceability, energy, and tempo of tracks by artist.
    - Identify the top 5 artists with the highest average track popularity.
3. **Data Transformation (3%)**
    - Create a new column called “energy_level” that classifies tracks as 'High Energy' (energy > 0.8) or 'Regular Energy' (energy ≤ 0.8).
    - Group the data by this new energy classification and calculate the average popularity and loudness for each energy_level.
4. **Data Exporting (2%)**
    - Export the data that have been classified as 'High Energy'.

Deliverables:

- A Jupyter notebook containing the Spark code with comments explaining each step.