Spotify Songs Analysis

Introduction

This Python data analysis project focuses on exploring and understanding a Spotify Songs dataset containing information about the songs available on Spotify. The dataset containing 30000 songs was obtained from Kaggle. The primary goal of this analysis is to identify any key characteristics that contribute to the song's popularity, such as danceability and genre. This project also aims to build a data model (regression) that predicts the popularity of a song.

Identify Key Characteristics that Contribute to Popularity

There is no clear correlation between popularity and the numeric characteristics of a song (danceability, liveness, tempo, acoustics, etc.). However, by categorizing the popularity into low (0-40), medium (40-80) and high (80-100) popularity, and comparing the distribution of the variables for each category, it is found that the **tempo distribution is becoming skewed to the left (between 75 – 100 beats per minute) from low to high popularity**. Besides, this analysis also found that certain genres and subgenres are **slightly more popular than others**, such as the **pop genre and post-teen pop subgenre**. The genre distribution is also found to differ by popularity level, such as **the proportion of EDM songs decreasing from low to high popularity,** at the same time **the proportion of Latin and Pop songs increases.**

Building a Regression Model to Predict a Song's Popularity

Linear Regression and K-Nearest Neighbors models are used to create the data model for predicting popularity based on predictors. Different scaling methods and combinations of predictors are tested. However, due to the **weak correlation between popularity and the predictors**, the performance of the models does not meet the desired accuracy, with the **highest R2 score 0.14376** (Linear Regression, predictors = variables with absolute correlation > 0.05). The project will continue to try with Classification model.