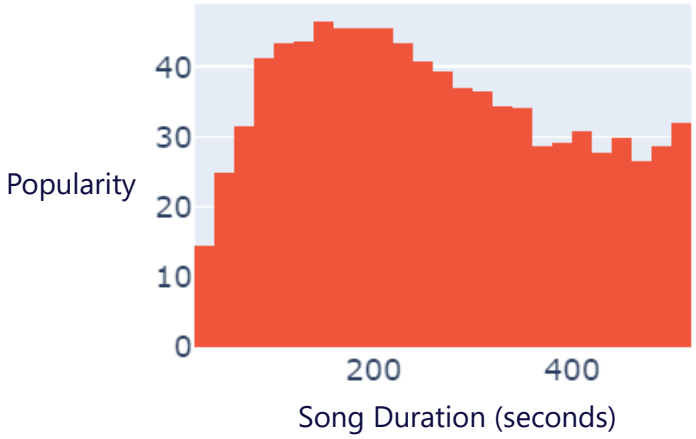
Spotify Songs Analysis

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, which contains 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 a song's popularity.

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, certain genres and subgenres are **slightly more popular than others**, such as **pop (genre) and post-teen pop (subgenre)**. There is also a pattern in the song duration in response to popularity, where the **songs between 65 and 340 seconds have popularity above average (~34)**:



By categorizing the popularity into low (0-39), medium (40-79) and high (80-100) popularity, and comparing the distribution of the variables for each category, there are some findings:

- The **tempo distribution is becoming skewed to the left (between 75 – 100 beats per minute) from low to high popularity**
- 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.**

The results are then visualized as a dashboard in Tableau.

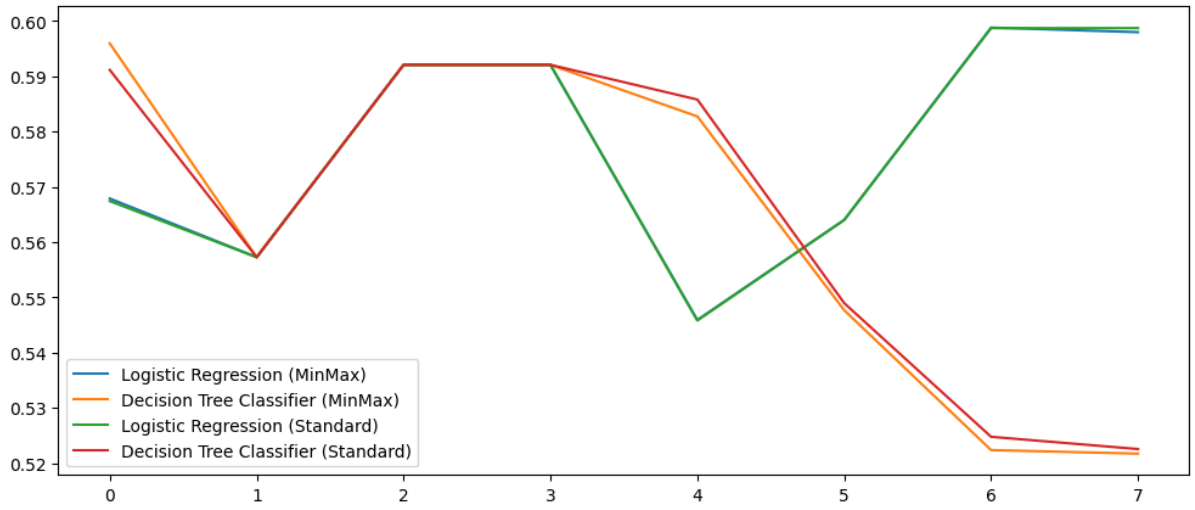
Building a Regression Model to Predict a Song's Popularity

Linear Regression and K-Nearest Neighbors models are used to create the data model for predicting popularity based on predictors. Different scaling methods *(MinMax & Standard)* 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 the Classification model.

Building a Classification Model to Predict a Song's Popularity (L/M/H)

Two methods are used: Logistic Regression & Decision Tree Classifier. The scaling methods and some of the predictor combinations used in the regression model are also used in the building of the classification model. The accuracy scores of each modelling and scaling method are stored in a Dict and then converted into a DataFrame. It is then visualized in a line chart for a better sight for observation and interpretation.

As the result, the **classification model outperforms the regression model**, with every modelling and scaling method achieving an accuracy score above 0.5. **The model of Logistic Regression with MinMaxScaler performs the best with predictor = duration + tempo + subgenres (accuracy score = 0.5988)**



0 – Numeric (danceability, tempo, key, valence, etc.)

1 – Genres

2 – Subgenres

3 – Genres + Subgenres

4 – Duration + Tempo

5 – Duration + Tempo + Genres

6 – Duration + Tempo + Subgenres

7 – Duration + Tempo + Genres + Subgenres