**A logo for college computing

Description automatically generated**

**Declaration**

By submitting this assessment, I confirm that I have read the CCT policy on academic misconduct and understand the implications of submitting work that is not my own or does not appropriately reference material taken from a third party or other source.

I declare it to be my own work and that all material from third parties has been appropriately referenced.

I further confirm that this work has not previously been submitted for assessment by myself or someone else in CCT College Dublin or any other higher education institution.

**Capstone Project Proposal**

**Unlocking the Charts: The Hidden Secrets From Hits**

**Submitted by: Luis Martins**

**Student ID: 2025053**

**Lecturer: Taufique Ahmed**

**Cohort: Higher Diploma Data Analytics Feb2025**

**Submission Date: 22nd Nov 2025**

Table of Contents

[**1. Introduction** 4](#_Toc213629134)

[**1.1. Background** 4](#_Toc213629135)

[**1.2. Objective of the Study** 4](#_Toc213629136)

[**1.3. Research Questions** 4](#_Toc213629137)

[**1.4. Scope and Limitations** 5](#_Toc213629138)

[**2. Methodology** 6](#_Toc213629139)

[**2.1. CRISP-DM Overview** 6](#_Toc213629140)

[**2.2. Data Preparation and Statistical Analysis Pipeline** 6](#_Toc213629141)

[**2.3. Model Development and Experimentation** 7](#_Toc213629142)

[**3. Data Description and Exploratory Data Analysis (EDA)** 8](#_Toc213629143)

[**3.1. Dataset Overview** 8](#_Toc213629144)

[**3.2. Data Sources and Ethical Considerations** 8](#_Toc213629145)

[**3.3. Descriptive Statistics** 9](#_Toc213629146)

[**3.4.1 Exploratory Data Analysis (EDA)** 9](#_Toc213629147)

[**3.4.2 Exploratory Data Analysis (EDA) – Results** 10](#_Toc213629148)

[Feature Correlation Heatmap 10](#_Toc213629149)

[10](#_Toc213629150)

[Tempo Distribution 11](#_Toc213629151)

[11](#_Toc213629152)

[Duration vs. Popularity 12](#_Toc213629153)

[Time Signature Analysis 13](#_Toc213629154)

[Top Artists by Song Count 14](#_Toc213629155)

[14](#_Toc213629156)

[Top Artists by Average Popularity 15](#_Toc213629157)

[15](#_Toc213629158)

[Artist Career Length 16](#_Toc213629159)

[16](#_Toc213629160)

[Top Subgenres by Popularity 17](#_Toc213629161)

[17](#_Toc213629162)

[Top Subgenres by Popularity 18](#_Toc213629163)

[18](#_Toc213629164)

[Album Release Frequency 19](#_Toc213629165)

[Popularity by Release Year 20](#_Toc213629166)

[Popularity Variance by Genre 21](#_Toc213629167)

[Seasonal Trends in Popularity 22](#_Toc213629168)

[22](#_Toc213629169)

[Trend of Popularity Over Time 23](#_Toc213629170) [23](#_Toc213629171)

[Top Playlists by Popularity 24](#_Toc213629172)

[24](#_Toc213629173)

[Musical Characteristics of Popular Playlists 25](#_Toc213629174)

[25](#_Toc213629175)

[**4. Data Preprocessing** 25](#_Toc213629176)

[**4.1. Normality Analysis and Scaling** 25](#_Toc213629177)

[**4.2. Dimensionality Reduction** 26](#_Toc213629178)

[**4.3. Class Imbalance** 27](#_Toc213629179)

[**4.4. Unsupervised Outlier Detection with "Isolation Forest"** 27](#_Toc213629180)

[**4.5. Multicollinearity and Feature Engineering** 28](#_Toc213629181)

[**4.6. Feature Importances & Explainability (SHAP, LIME)** 29](#_Toc213629182)

[**5. Statistical Tests Summary** 30](#_Toc213629183)

[**5.1. Parametric Tests** 30](#_Toc213629184)

[**5.2. Non-Parametric Tests** 31](#_Toc213629185)

[**6. Model Selection and Training** 32](#_Toc213629186)

[**6.1. Model Categories and Algorithms (Classifier and Regressor)** 32](#_Toc213629187)

[**6.2. Evaluation Metrics** 32](#_Toc213629188)

[**6.3. Regression vs. Binary Classification Strategy** 33](#_Toc213629189)

[**7. Experiments and Hyperparameter Implementation** 33](#_Toc213629190)

[**7.1. Experiment 1: Binary Classification Models** 35](#_Toc213629191)

[**7.2. Experiment 2: Regression Models** 35](#_Toc213629192)

[**7.3. Experiment 3: Feature Importance & Explainability** 36](#_Toc213629193)

[**8. Cross-Experiment Summary** 37](#_Toc213629194)

[**8.1. Binary Classification Top Results** 37](#_Toc213629195)

[**9. Model Evaluation** 38](#_Toc213629196)

[**9.1. Best Models Across All Experiments** 38](#_Toc213629197)

[**9.1.1 Best Classifier Algorithm** 38](#_Toc213629198)

[**9.1.2 Best Regressor Algorithm** 39](#_Toc213629199)

[**9.2. Recommendation** 40](#_Toc213629200)

[**9.3. Practical Recommendations** 41](#_Toc213629201)

[**9.4. Challenges** 41](#_Toc213629202)

[**9.5. Future Work** 42](#_Toc213629203)

[**Use of Artificial Intelligence:** 42](#_Toc213629204)

[**References List (Harvard Style)** 42](#_Toc213629205)

## **1. Introduction**

### **1.1. Background**

Spotify, the largest and most influential digital music distribution platform today has transformed the way we listen to and create music. The sheer number of artists creating content and competing for the same listener is overwhelming. Artists are struggling to find ways to be noticed in a crowded space which means finding ways to get heard is becoming increasingly difficult – it is called “discovery”. (Porter, 2008).

In addition to being technical in nature the discovery problem is strategic in nature. To gain a competitive advantage in this complex digital environment understanding the inner workings of this ecosystem is crucial. (Davenport and Harris, 2007). Success of a song today is not just about the quality of the music or the creativity of the artist. It is about how well the song fits into the mechanics of Spotify and the preferences of Spotify’s listeners. Analyzing the characteristics of successful songs will provide a data-driven method to develop a strategy for success and will move away from intuition and towards using data to make decisions.

### **1.2. Objective of the Study**

The purpose of this research is to determine and analyze the differences between the audio and metadata features of highly popular songs and those of less popular songs on Spotify.

This study seeks to go beyond simple correlations and provide practical advice to artists and producers. This research will seek to provide a set of strategic recommendations that can be used to guide artists and producers to make better decisions during the production and release of their music to increase listenership and increase their chances of success. (Johnson and Scholes, 1993). This objective is a direct application of the use of data to inform future actions (Vaughan, 2020).

### **1.3. Research Questions**

To achieve the stated objective, this study will address the following key research questions:

1. What is the statistical relationship between core audio features (e.g., danceability, energy, loudness, valence) and a song's classification as "popular"?
2. Do structural elements of a song, such as duration, tempo, key, and mode, show a significant correlation with its popularity?
3. To what extent do temporal factors, specifically the release\_year, influence a song's popularity within the dataset?
4. Can a machine learning model be developed to accurately predict whether a song will achieve high popularity based on its intrinsic features?

### 

### **1.4. Scope and Limitations**

The scope and limitations of this study have some methodological constraints:

**Scope:**

* The study only examines data from the Spotify platform and utilized a single static, consolidated dataset obtained from Kaggle.
* The project employed quantitative methodologies (Exploratory Data Analysis (EDA), Statistical Testing and Classification-Based Machine Learning (Random Forest, SVM)) in order to analyze the relationship between the music characteristics of popular songs and their actual popularity.

"Popularity" was defined as a binary classification (1 = High, 0 = Low) based on how the source dataset defines these classifications.

**Limitations:**

* **Data Source:** The Kaggle dataset used in this study is a static export and not a live feed from the Spotify API; therefore, the definitions of "High" and "Low Popularity" are predetermined by the dataset and do not necessarily reflect the entire range of what constitutes success.
* **External Variables:** This study is limited to analyzing intrinsic audio and metadata features. It **does not account for crucial external factors** that heavily influence popularity, such as marketing budgets, social media promotion, artist brand recognition, or algorithmic playlist placement.
* **Correlation vs. Causation:** The results of this study will show statistically significant correlations. However, it is very important to avoid interpreting the existence of these correlations as a causal relationship between the characteristic(s) and the song becoming popular. For example, if there is a correlation between a certain tempo and popular songs, it does not mean that the tempo itself causes the song to be popular.

## **2. Methodology**

This chapter details the structured approach undertaken to answer the research questions. The methodology is guided by the **Cross-Industry Standard Process for Data Mining (CRISP-DM)**, ensuring a robust and repeatable analytical process. This section outlines the project's phases, the specific data preparation and analysis pipeline, and the strategy for model development and experimentation.

### **2.1. CRISP-DM Overview**

CRISP-DM is a systematic approach to data driven projects. It has six major phases to guide your work:

1. **Business Understanding:** The business problem we are trying to solve with our data is music discovery in a crowded digital marketplace. Our objective is to find ways to help artists (organization) use data to have a competitive advantage to know what makes popular songs.
2. **Data Understanding:** We sourced two .CSV files from kaggle, explored the feature sets in the file to know what we had and where we could expect to find popular and unpopular songs based on the distributions of those features. **Total records: 4829; Total features: 25**
3. **Data Preparation:** The most time-consuming pre-analytic phase, covered in section 2.2, involved cleaning, merging and preparing the raw data for analysis and modeling.
4. **Modeling:** Section 2.3 covers the modeling phase. It is here that we select and apply several machine learning methods to the prepared data to predict song popularity.
5. **Evaluation:** Each of these models will be evaluated to see how well they can predict the popularity of songs and more importantly to see if they can provide us with some meaningful insights related to the research questions.
6. **Deployment:** In the case of this academic project, deployment means to take the results from the last model and turn them into actionable insights and strategic recommendations for the final report.

### **2.2. Data Preparation and Statistical Analysis Pipeline**

It was built a complete data preparation pipeline to ensure the integrity and quality of the data necessary for meaningful analysis. (Reis and Housley, 2022).

* **Data Consolidation:** Two source files (high\_popularity\_spotify\_data.csv and low\_popularity\_spotify\_data.csv) were imported. A binary target variable, popularity, was created (1 for high, 0 for low) so that the research question can be transformed into a binary classification problem. The two datasets were then merged into a single master DataFrame.
* **Data Cleaning:** To clean the data for quality purposes, two cleaning tasks were performed. First, all rows that contained NaN (null) values were eliminated. Then, duplicate rows were identified and removed to avoid biasing the analysis.
* **Feature Selection and Engineering:**
  + **Selection:** Columns irrelevant to the modeling objective, such as track\_id, track\_href, and analysis\_url, were dropped.
  + **Engineering:** The track\_album\_release\_date column, a string object, was engineered. The **release\_year**, **release\_month**, and **release\_day** were extracted as separate numerical features, enabling temporal analysis.
* **Data Transformation:** Audio features such as key and mode were changed from numeric to categorical data types. MinMaxScaler was applied to the main numeric audio features (danceability, energy, loudness, tempo). Normalizing is important because it puts all of the features on the same scale (0 to 1) so that algorithms that rely on the magnitude of features (such as Support Vector Machines) can operate at their best.
* **Statistical Analysis:** EDA was used to explore the prepared data to look at the relationships between individual features and the popularity target. This is a direct way to address the research questions.

### **2.3. Model Development and Experimentation**

The core of the predictive methodology is developing and testing machine learning models to find the most important factors that drive popularity.

* **Problem Formulation:** The task is formulated as a supervised binary classification problem. The model is trained on the prepared feature set (X) to predict the binary popularity target (y).
* **Train-Test Split:** The combined dataset was split into a training set and a testing set (for example an 80/20 split). The models are trained on the training data and their performance is tested on the unseen test data to determine their generalizability in the real world.
* **Algorithm Selection:** A suite of robust classification algorithms was selected, as identified in the project notebook:
  + **Random Forest Classifier:** An ensemble method chosen for its high accuracy, robustness against overfitting, and its intrinsic ability to calculate feature importance.
  + **Support Vector Classifier (SVC):** A powerful algorithm effective in high-dimensional spaces, suitable for finding a clear margin of separation between the two popularity classes.
  + **XGBoost:** A gradient-boosting algorithm renowned for its high performance, speed, and predictive power.
* **Model Evaluation:** The performance of each model will be assessed using a confusion matrix and common classification metrics such as Accuracy, Precision, Recall and F1-Score. These metrics will give a complete picture of how well the model predicts both popular and non-popular songs.
* **Feature Importance Analysis:** One of the goals of this project is to go beyond simple prediction. The feature importance scores generated by the Random Forest and XGBoost models will be extracted and analyzed. This final step is crucial because it directly gives the "actionable insights" to quantify exactly which specific audio features (e.g., energy, danceability) are the strongest predictors of a song's success.

## **3. Data Description and Exploratory Data Analysis (EDA)**

Exploring the data was an important part of the research because it gave us a better understanding of what we had at our disposal and helped us to develop some preliminary ideas of the types of analyses we could perform. Data descriptions and exploratory data analysis (EDA) form the foundation for the remainder of the modeling in this study.

### 

### **3.1. Dataset Overview**

This study utilizes a large dataset of songs from Spotify (Gupta and Mamta, 2023). The data came in the form of two files representing “High Popularity” and “Low Popularity” tracks. To prepare the data for analysis, the two files were combined and a binary target variable was created (i.e., a “high popularity” track was assigned a 1 and a “low popularity” track was assigned a 0). The data contained a diverse set of variables describing each song which are summarized into two general groups of feature types;

The dataset contains a rich set of features for each song, which can be grouped into two main categories:

1. **Audio Features (Quantitative):** These are quantitative, objective, numerical values given by Spotify that describe the song’s musical content. Some examples of key features are danceability, energy, loudness, valence (positive musical characteristics), tempo (beats per minute), and duration\_ms.
2. **Metadata & Categorical Features:** This group includes descriptors of the song.
   * **Categorical:** key (e.g., C, C#) and mode (Major/Minor).
   * **Temporal:** track\_album\_release\_date, which was engineered to extract release\_year, release\_month, and release\_day as separate, analyzable features.

Irrelevant identifiers for this analysis, such as track\_id and track\_href, were removed during the data preparation phase to streamline the dataset for modeling.

### **3.2. Data Sources and Ethical Considerations**

**Data Source:** The data used in this project was obtained from a public access dataset on the Kaggle data science platform (Dataset: "Spotify Music Dataset"). The use of a public access dataset is ideal for academic studies due to its transparency and ability to provide the same results as reported here.

**Ethical Considerations:** The use of this data has been deemed ethically acceptable for several reasons.

* **Anonymity:** The data in question describes music tracks and not individual people. There is no personally identifiable information (PII) regarding Spotify users or their listening habits in the data.
* **Purpose:** The data is being used solely for academic purposes to identify the broader trends present in music and consistent with the intent of such publicly accessible data.
* **Bias Awareness:** One of the most significant ethical and methodological concerns is the pre-existing bias in the “popularity” label of the dataset. The criteria for labeling a track as either “high” or “low” popularity are unknown and do not necessarily reflect all aspects of musical success. As stated earlier, the study’s conclusions should be viewed as specific to this particular definition of popularity.

### **3.3. Descriptive Statistics**

As a fundamental requirement to understand the dataset, descriptive statistics were produced to summarize both the central tendencies and the distribution of the data. This process provided an appreciation of the “typical” song and the amount of variation present in the data.

* **Continuous Variables:** Measures of central tendency (mean, median) and dispersion (standard deviation, IQR) were compared for the continuous features (e.g., danceability and energy) to help answer questions like “What is the typical loudness of a song and how much will it usually vary?”
* **Categorical Variables:** Frequency distributions were developed to measure the proportion of each category (e.g., “What percent of the songs in the dataset are written in a major key?”) for the categorical features (e.g., key and mode).
* These descriptive statistics represented a basic level of understanding of the structure of the data prior to visual exploration and hypothesis testing.

### **3.4.1 Exploratory Data Analysis (EDA)**

EDA was conducted to visually examine the data and identify possible relationships among the features. This phase of investigation is critical to identifying possible evidence of the features that relate to the target variable, popularity.

The EDA process involved:

1. **Correlation Analysis:** A correlation matrix was developed to analyze the relationship between the principal audio features (e.g., energy, danceability, valence, loudness) and the popularity variable. The goal of this was to identify those features that exhibited the largest linear relationship with the target variable.
2. **Distribution Analysis:** Histograms and box plots were developed to visualize the differences in the distributions of certain features (e.g., tempo, duration\_ms) in the two popularity levels. This facilitated comparisons of whether popular songs are typically faster or longer than unpopular ones.
3. **Categorical Investigation:** The examination of categorical features (e.g., key and mode) was extended to determine if popular songs exhibit a preference to be composed in certain keys or modes.
4. **Hypothesis Formulation:** The findings from the EDA (e.g., there may exist a difference in length between popular and unpopular songs) were used to generate specific hypotheses that were subsequently tested using formal statistical procedures (e.g., t-test) as documented in the project notebook.

### **3.4.2 Exploratory Data Analysis (EDA) – Results**

### Feature Correlation Heatmap

### A diagram of a number of different colored squares AI-generated content may be incorrect.

Figure 1: Correlation of Audio Features with Track Popularity

This correlation chart displays the connection between popular songs' primary audio attributes and their popularity in a graphical format. The color gradient ranging from 0.2 (blue), representing little to no positive correlation, to 1.0 (red), representing strong positive correlation, represents the strength of the positive linear correlation among the variables.

This chart is quite relevant as it indicates that while the relationship between the various audio attributes and popularity is relatively weak (loudness=0.22; energy=0.19) the most significant correlations are with those two attributes (loudness and energy). Additionally, the attribute with the least correlation to popularity is valence (the musical happiness). This data provides insight into answering your first research question and provides evidence that the attributes of loudness and energy will need to be included in your model.

### Tempo Distribution

## **A graph of different colored bars AI-generated content may be incorrect.**

Figure 2: Tempo (BPM) Distribution for Popular vs. Less Popular Songs

The histograms below compare the tempos (in Beats Per Minute, or BPM) of songs within the "Popular" group (blue) and the "Less Popular" group (red).

Relevance and Conclusion: This is related to song structure analysis, because it appears that tempos between 120-130 BPM are most common with popular tracks (as they are the most dense on the graph). In contrast, the less popular songs are more distributed over the range of tempos, indicating that a tempo that falls into the "popular sweet spot" may be an important characteristic of a successful track.

### Duration vs. Popularity

A graph with blue and white lines

AI-generated content may be incorrect.

Figure 3: Song Duration Distribution by Popularity Score

This box plot depicts the distribution of the duration of songs in terms of their level of popularity (from 76 to 100).

Relevance: The data indicates that the most consistently represented median duration for most popular songs lies between 200,000-250,000 milliseconds (which is approximately 3 minutes 20 seconds to 4 minutes 10 seconds). It does not appear there are any trends indicating more or less popular songs have greater durations, however it also supports the fact that most hit songs are all relatively similar in length to what would be considered standard in a radio format.

### Time Signature Analysis

A comparison of a bar graph

AI-generated content may be incorrect.

Figure 4: Impact of Time Signature on Danceability and Popularity

The two bar charts show the Time\_Signature data. The bar on the left side displays the average Danceability per Time\_Signature. The bar on the right side of the chart shows the average Popularity.

Summary & Importance: That is one of the clearest and most important findings from this study. The Time\_Signature with an average of 4 (the 4/4 rhythm which is the most common time in pop music) has both the highest average danceability and the highest average popularity. This indicates that songs written in 4/4 time will have a higher probability of being popular and danceable.

### Top Artists by Song Count

## **A graph of music artists with different colors AI-generated content may be incorrect.**

Figure 5: Top 10 Artists by Number of Popular Songs in Dataset

This is a bar graph listing all 10 artists with the most songs in the "popular" count of the database.

This will help describe our dataset by showing that artists such as Sabrina Carpenter and Linkin Park represent the largest portion of the "popular" section of our "popular" category of songs. It could be said to provide a descriptive piece of the data but it does not offer a piece for analyzing the reasons behind the popularity of each song.

### Top Artists by Average Popularity

## **A graph of music artists AI-generated content may be incorrect.**

Figure 6: Top 10 Artists by Average Song Popularity Score

This bar graph is ordered from greatest to least based on the artist with the most (or count) of their songs listed as "popular" in this data set.

This graph will allow us to see how many times each artist appears in the "popular" category and help us describe our data. We can also determine that we have frequent appearances by Sabrina Carpenter and Linkin Park in the "popular" category. The purpose of this graph is to describe the data rather than analyze why certain artists' songs are classified as "popular".

### Artist Career Length

## **A graph of a number of dots AI-generated content may be incorrect.**

Figure 7: Relationship Between Artist Career Length and Average Song Popularity

Summary of Scatter Plot and Relationship between Career Length and Average Popularity of Artists

Findings/Relationships: The graph above does show that there is no obvious correlation between an artist’s "Career Length (in Years)" and their "Average Popularity". There are artists with the same level of popularity who have had careers of only one year, as well as artists with the same level of popularity who have had long lasting careers of 30 years. This indicates that both new and established artists are able to create “hit” songs.

### Top Subgenres by Popularity

## **A graph of different colored lines AI-generated content may be incorrect.**

Figure 8: Audio Feature 'Fingerprints' by Genre

The above bar chart provides a detailed view of the average “fingerprint” for multiple genres. The chart plots the scaled average of music features such as energy, danceability, valence (positive), tempo, loudness, and acousticness for each genre.

Relevance & Conclusion: This is a rich bar chart. Each genre has a distinct audio profile. For example, j-pop and k-pop have a high amount of energy, danceability, and loudness. Ambient and folk are the exact opposite of those two genres as they contain a high level of acousticness and a lower level of energy. As this is a visual representation of the differences between genres in terms of their sonic attributes; it supports the notion that the rules governing what makes something popular may differ from one genre to another.

### Top Subgenres by Popularity

## **A graph of a number of people AI-generated content may be incorrect.**

Figure 9: Top 10 Subgenres by Average Popularity

Reason for Bar Chart: This bar chart lists the top 10 sub-genres by their average "popularity" score.

Relevance and Conclusion: This graph provides evidence to identify which sub-categories of songs have the highest popularity in this data set. The "global", "mainstream" genres being at the top of the list makes sense as they are among the most well known genres. This also will help to define a "popular" sounding song as it will typically be "mainstream", "soft", "melodic".

### Album Release Frequency

A graph of a music album release

AI-generated content may be incorrect.

Figure 10: Artist's Total Albums Released vs. Average Song Popularity

Explanation: The Scatter Plot in Figure 8 illustrates how the total number of albums produced by a single artist relate to their average popularity.

Relevance & Conclusion: As with Figure 7 "Career Length", there is no apparent relationship between "Quantity" (number of albums) and "Quality" (average popularity). Only approximately 2% of all artists have released more than 3 albums. An artist can produce very few albums and still be very popular or they can produce many albums and yet be relatively unpopular. Therefore, the amount of an artist's album production may not determine the quality of those productions.

### Popularity by Release Year

A graph showing different colored lines

AI-generated content may be incorrect.

Figure 11: Distribution of Track Popularity by Release Year

Summary: The Box Plot is showing how track popularity is distributed by all years of "Release Year" from 1957 to 2024.

Concluding Statement: The findings shown here are fundamental and will provide direct support to your hypothesis or research question. The findings show classic hits do exist for each year from 1960 through 1980. However, the median value of popularity (line inside the box), and the overall popularity are both most concentrated between the late 1990s and the late 2010s. These results will confirm release\_year as a valuable feature for your modeling process.

### Popularity Variance by Genre

A graph of a number of people

AI-generated content may be incorrect.

Figure 12: Variance of Popularity Scores Within Each Genre

Explanation: The bar graph measures the “Variability” of each genre. Variability is measured in terms of how much the individual song ratings vary from one another.

Conclusion & Relevance: There is some interesting information here. For example, genres such as pop, latin and indian have variability, which means there may be some extremely popular and unappealing songs. On the other hand, those genres that have little to no variability (such as wellness and r&b) are relatively consistent. This is a good insight into the potential usefulness of including the genre variable in your model to better understand why the distribution of ratings differs across genres.

### Seasonal Trends in Popularity

## **A graph of different colored bars AI-generated content may be incorrect.**

Figure 13: Average Song Popularity by Release Month

The graph displays the “Average Popularity” for each song, by month it was created/released.

Reasoning/Relevance: The purpose of this graph is to identify whether there are any seasonal patterns. What the data does show is that while very subtle; the average popularity of all songs (that were created and/or released during the month of January) is the highest. Also, the average popularity of songs (released during June – October) has an upward trend over the course of the summer and autumn months. This further supports why you should create a feature out of the engineering of release\_month.

### Trend of Popularity Over Time

## **A graph showing the trend of song popularity over the years AI-generated content may be incorrect.**

Figure 14: Average Song Popularity by Release Year (1957-2024)

This line graph illustrates the “average popularity” of one score per “release year,” illustrating the historical pattern.

This chart (related to Fig 11) is very clear: Average popularity was high during the years from 1960–1980, fell rapidly after 1990; it has slowly declined with some irregularities since then. This is a good example of how release\_year is an important variable that reflects a dominant trend in the data**.**

### Top Playlists by Popularity

## **A graph of a number of playlists AI-generated content may be incorrect.**

Figure 15: Top 10 Playlists by Average Song Popularity

Explanation: This bar chart ranks the "Top 10 Playlists" in the dataset based on the average popularity of the songs they contain.

Conclusion & Relevance: This chart helps us understand the *source* of popular music in the data. Playlists like "Global Top 50" and "Today's Top Hits" are, as expected, filled with the most popular songs. This confirms that these playlists are a good proxy for "what is popular."

### Musical Characteristics of Popular Playlists

## **A graph of different colored bars AI-generated content may be incorrect.**

Figure 16: Audio Feature 'Fingerprints' of Top 10 Playlists

Similarities: This chart is similar to figure 8, but this chart illustrates the audio fingerprint for the "top ten playlists" as illustrated in the prior chart.

Conclusion/Relevance: This is the most important chart. The chart presents an analysis of the sounds of "hit" songs. As indicated by the top playlists (global top 50, today's top hits, top gaming tracks), each playlist has a very similar pattern: high scaled energy, high scaled danceability, high scaled loudness and low scaled acousticness. This is a powerful conclusion that will help you support the "practical recommendations" in your final report.

## **4. Data Preprocessing**

The 4th Chapter of this report provides a detailed account of the processes that were followed for pre-processing of the collected datasets post exploratory data analysis (EDA). The Data Preprocessing Phase encompasses cleaning and preparing the data; enhancing the quality of the features; and preparing the data for use in high performance machine learning. This aligns with the Data Preparation phase of the CRISP-DM Framework.

### **4.1. Normality Analysis and Scaling**

Although there was no emphasis on formally testing for normality (i.e. Shapiro-Wilk test), an examination of the distributions of the various audio features during the EDA indicated that the vast majority of these did not conform to a standard normal distribution.

A group of graphs showing different colored lines

AI-generated content may be incorrect.

Figure 17: Distributions of Key Numerical Features

What is perhaps more important is that models such as Support Vector Machines (SVM) are sensitive to the scales of the input features. For example, the MinMaxScaler was used to normalize all the numeric audio features (e.g. danceability, energy, loudness). As stated previously, the MinMaxScaler transforms each feature to a common range [0,1]. This allows no single feature to have dominance over the model’s learning process due to a larger magnitude. More importantly, this normalization is necessary for preventing bias in the model, and for helping the model converge.

### **4.2. Dimensionality Reduction**

Dimensionality reduction occurred via manual feature selection and not through automated methods such as Principal Component Analysis (PCA). As per the business requirements, multiple columns were identified as being non-relevant for predicting popularity, and therefore they were removed from the dataset.

These columns (e.g., track\_id, track\_href, analysis\_url, track\_name, artist\_name) serve as identifiers or qualitative text fields that, while useful for human readability purposes, do not contain any predictive value for a quantitative model and will add significant amounts of noise. By intentionally reducing the number of dimensions in the dataset, we are able to significantly reduce the amount of computation required, and focus the models on a curated set of relevant features.

### **4.3. Class Imbalance**

In addition to assessing the target variable (popularity) of the classification problem, another critical step in developing a classification model is determining whether the classes of the target variable are significantly imbalanced.

The dataset used in this study was developed using a single file containing high-popularity tracks (Class 1) and a second file containing low-popularity tracks (Class 0).

A screenshot of a computer code

AI-generated content may be incorrect.

Figure 18:Distribution of Popularity Categories

Prior to developing the classification model, an examination of the value counts of the track\_popularity variable was conducted to assess the level of imbalance present. An extreme imbalance between two classes can lead to biased results for the model, resulting in the model producing predictions of the majority class at a disproportionate rate.

### **4.4. Unsupervised Outlier Detection with "Isolation Forest"**

In this project, the Interquartile Range (IQR) method was used to identify and remove statistical outliers from the dataset. This was an important step in the data cleaning process to ensure the quality of the data before modeling.

The method was applied to each key numeric feature (like danceability, energy, tempo, etc.) individually. For every feature, the code first calculated the 25th percentile (Q1) and the 75th percentile (Q3). The IQR was then found by subtracting Q1 from Q3.

Using this IQR value, a valid "acceptable" range was defined. The lower boundary was set at Q1 - 1.5 \* IQR, and the upper boundary was set at Q3 + 1.5 \* IQR. Any data point (row) with a value that fell outside this range for that specific feature was classified as an outlier and removed from the dataset. This process was repeated for all selected features, cleaning the data one feature at a time.

A screenshot of a computer code

AI-generated content may be incorrect.

Figure 19: Count of Outliers Detected per Feature (IQR Method)

### **4.5. Multicollinearity and Feature Engineering**

* This phase involved both the creation of new features and the assessment of relationships between existing ones.
* **Feature Engineering:** The most significant engineering step was the parsing of the track\_album\_release\_date column. This single string object was transformed into three new numerical features: release\_year, release\_month, and release\_day. This is a strategically vital step, as it "unlocks" the temporal dimension of the data, allowing the model to analyze trends and determine if a song's release date has any bearing on its eventual popularity.
* **Multicollinearity Assessment:** Multicollinearity occurs when predictor variables are highly correlated with each other, which can destabilize model coefficients and reduce the interpretability of feature importance. The correlation matrix generated during the EDA phase (Chapter 3) was used to assess this. High correlations were observed (between energy and loudness), but no features were removed, as tree-based models (like Random Forest and XGBoost) are generally robust to multicollinearity..
  + To ensure the model is reliable, it is important to check if the independent variables are highly correlated, a problem known as multicollinearity. To do this, the Variance Inflation Factor (VIF) was calculated for each numeric audio feature.
  + The VIF method measures how much a single feature is "inflated" or explained by all the other features in the dataset. The code created a list of VIF scores for each variable. A high VIF score (often above 5 or 10) indicates that a feature is strongly correlated with others. This analysis is crucial for identifying which features might need to be removed to build a more stable and accurate predictive model.

A graph of a number of bars

AI-generated content may be incorrect.

Figure 20: Variance Inflation Factor (VIF) for Numeric Features

### **4.6. Feature Importances & Explainability (SHAP, LIME)**

While this is primarily an evaluation strategy, it is fundamentally linked to the preprocessing and modeling goals. The objective of this study is not just to predict popularity, but to explain it.

To achieve this, the following tools are designated for the model evaluation phase:

* **RF Feature Importances:** The RandomForestClassifier and XGBoost models both possess an intrinsic feature\_importances\_ attribute. This will be the primary tool used to rank features and identify the key drivers of popularity (e.s., "is danceability more important than tempo?").
* **Explainability Tools (SHAP/LIME):** These tools are noted as the next step for deeper analysis. While feature\_importances\_ gives a global overview, tools like SHAP (SHapley Additive exPlanations) would provide a more granular, instance-level explanation, showing how the model "thinks" for each individual song prediction. This directly serves the project's goal of creating "actionable insights." XGBRegressor was chosen for this analysis.

A graph of different values

AI-generated content may be incorrect.

Figure 21: SHAP Summary Plot for Feature Importance and Impact

## **5. Statistical Tests Summary**

After conducting the first Exploratory Data Analysis (EDA), a set of formal statistical tests were conducted as part of this phase. These tests were to take the preliminary visual identification to a quantitative level of verification and determine if there are statistically significant differences in the relationships that were visually identified or merely due to chance(Indeed Career Guide, n.d.).

The tests for this section of analysis will be divided into two categories: parametric and non-parametric tests; each test has been selected on the basis of the nature of the data (i.e., continuous vs. categorical) and the extent to which it fulfills the assumptions of that specific test.

### **5.1. Parametric Tests**

Parametric tests assume that your data follows a particular shape, usually that it is normally distributed. With a very large sample size, we can invoke the Central Limit Theorem and assume that the sampling distribution of the mean will be approximately normal, and therefore allow us to confidently apply parametric tests.

**Independent Samples t-test:** The primary parametric test employed was the **Independent Samples t-test**.

* **Objective:** This test was used to compare the means of continuous audio features between the two independent groups: "high popularity" (Class 1) and "low popularity" (Class 0).
* **Application:** It was applied to key quantitative variables identified in the EDA, such as:
  + duration\_ms
  + danceability
  + energy
  + loudness
  + tempo
* **Hypothesis:** For each feature, the null hypothesis (H₀) stated that there is no difference in the mean of that feature between the two popularity groups. The alternative hypothesis (H₁) stated that there is a significant difference in the mean of that feature between the two popularity groups.
* **Outcome:** The p-value obtained from each test represents the probability of observing the data if the null hypothesis were true. A small p-value (usually less than 0.05) will indicate that the null hypothesis should be rejected, suggesting that the feature in question has a significantly larger (or smaller) average value in popular songs compared to unpopular songs.

### **5.2. Non-Parametric Tests**

Non-parametric tests do not require knowledge of how the data is shaped. Therefore, they were chosen for analysis of relationships between categorical variables, since mean-based tests cannot be used for such analysis.

**Chi-Square (χ²) Test of Independence:** The principal non-parametric test used was the **Chi-Square (χ²) Test of Independence (Kateri and Kateri, 2024)**.

* **Objective:** The purpose of this test is to find out if there is a significant association (dependency) between two categorical variables.
* **Application:** The Chi-Square test was applied to examine the relationship between the binary target variable (popularity) and the categorical audio characteristics:
  + mode (Major vs. Minor)
  + key (C, C#, D, etc.)
* **Hypothesis:** The null hypothesis (H₀) stated that the popularity of a song is independent of its mode (and/or key). The alternative hypothesis (H₁) stated that there is a statistically significant association between these two characteristics.
* **Outcome:** If the Chi-Square test results in a low p-value, then the observed frequency of popular songs will be found to be significantly different between different keys or modes, indicating that there is a predictive relationship between a song's success and its mode or key.

## **6. Model Selection and Training**

In order to address the capstone project's objective of establishing why a song becomes popular, a dual-model strategy is employed which incorporates both binary classification and regression. In this section, the models chosen for each model type, the measures of success for them, and the rationale for employing a multi-faceted methodology is discussed.

### **6.1. Model Categories and Algorithms (Classifier and Regressor)**

The project utilizes two separate model types to provide an inclusive view of what constitutes "popular":

1. Binary Classification: The binary classifier determines whether a song will be classified as "Popular" (1) or "Not Popular" (0), utilizing the high and low popularity derived popularity characteristic. The results of this classifier are a definitive direction for an artist or producer.

2. Regression: The regressor determines the exact track\_popularity score (a continuous score from 0-100). Utilizing this model allows for a more precise determination of the degree to which a song will likely achieve popularity.

Robust ensemble methods, in addition to powerful kernel-based models, were employed for both model types (FrancescaLazzeri, 2022).. Additionally, all models were trained with GridSearchCV to ensure optimal hyperparameter selection, and StandardScaler was utilized to scale the features and normalize their influence.

The particular algorithms employed for each model type are:

* **Classifier Algorithms:**
  + Random Forest Classifier
  + Support Vector Classifier (SVC)
  + Gradient Boosting Classifier
* **Regressor Algorithms:**
  + XGBoost Regressor
  + Support Vector Regressor (SVR)
  + Gradient Boosting Regressor

### **6.2. Evaluation Metrics**

The measures of success for each model were determined by the distinct performance metrics employed for classification and regression tasks, as shown in the model training code.

**Classification Metrics**

• Accuracy: This metric evaluates the total number of accurate classifications made by the model (i.e., (True Positives + True Negatives) / Total Population). Accuracy provides a basic understanding of the model's overall correctness.

• Cross-Validation (CV) Accuracy: To ensure the model has not over-fit the training data, 5-Fold Cross Validation accuracy was also evaluated. CV Accuracy provides a more reliable assessment of the model's ability to predict the outcomes for unknown data.

**Regression Metrics**

• **R-squared (R2) Score**: The R2 Score measures the amount of variance in the track\_popularity score that is explained by the audio features. An R2 Score close to 1 indicates that the model is able to explain a significant portion of the variance associated with popularity.

• **Root Mean Squared Error (RMSE)**: This metric represents the standard deviation of the residuals (prediction error). RMSE is measured in the same units as the target variable (popularity score) and is indicative of the average distance from the actual popularity score that the model's prediction is from.

• **Cross-Validation (CV) R2 Score**: Similarly to the classification models, a 5-Fold Cross Validated R2 Score was assessed to determine the model's ability to generalize its predictive power to future, unknown data.

### **6.3. Regression vs. Binary Classification Strategy**

A primary component of the project's analytical strategy is the decision to utilize both classification and regression models.

The binary classification model answers the high-level business question: "Will this new song be a hit?" The binary classification model trains upon the popularity feature (1/0), providing actionable, "yes/no", "go/no-go" recommendations for artists and producers and assisting in identifying the characteristics of successful songs.

The regression model answers a more detailed question: "How big of a hit will this be?" The regression model is able to make a more precise prediction of how well a song will do based upon its continuous track\_popularity score. The model can assist in differentiating between a song that is likely to be somewhat popular (e.g., a score of 60) and a major hit (e.g., a score of 90).

Through the combination of both the classification and regression models, the project offers a more robust set of predictive tools; the classification models offer strategic guidance (strategic decisions) and the regression models assist in setting expectations (managing expectations).

## **7. Experiments and Hyperparameter Implementation**

This chapter presents the execution and results of the machine learning experiments. Following the methodology defined in Chapter 6, this section is structured into three distinct experiments:

1. A primary experiment using **binary classification** to predict song popularity.
2. A secondary experiment using **regression** to assess an alternative predictive strategy.
3. A final analysis focused on **feature importance** to extract strategic insights.

In order to go beyond the baseline performance of the models, we have used the GridSearchCV function from scikit-learn to perform hyperparameter tuning. The GridSearchCV function performs a complete search of all combinations of parameters in a given “grid” for a specific algorithm.

**Cross-validation of Three Folds**

All experiments were performed with a 3-fold cross-validation (cv = 3). Each time the GridSearchCV is run, it will retrain the model on a subset of the training data and calculate a score for that particular subset of the training data. Since cross-validation is performed by default during the grid search for all experiments, this ensures that the best set of parameters for each model is robust and minimizes overfitting.

Parameter Grids: It was defined the following parameter grids to be tested for each model:

**Classification Model Grids:**

* **RandomForestClassifier (rf\_params):**
  + 'n\_estimators': [100, 150]
  + 'max\_depth': [10, 20]
  + 'min\_samples\_split': [2, 5]
* **Support Vector Classifier (svc\_params):**
  + 'C': [0.1, 1, 10]
  + 'kernel': ['rbf', 'linear']
* **GradientBoostingClassifier (gbc\_params):**
  + 'n\_estimators': [100, 150]
  + 'learning\_rate': [0.05, 0.1]
  + 'max\_depth': [3, 5]

**Regression Model Grids:**

* **XGBoostRegressor (xgb\_params):**
  + 'max\_depth': [7, 15]
  + 'learning\_rate': [0.01, 0.1]
  + 'n\_estimators': [100, 300]
* **Support Vector Regressor (svr\_params):**
  + 'C': [0.1, 10, 100]
  + 'kernel': ['rbf', 'linear']
  + 'epsilon': [0.01, 0.05, 0.1, 0.2, 0.5]
* **GradientBoostingRegressor (gbr\_params):**
  + 'n\_estimators': [100, 150]
  + 'learning\_rate': [0.05, 0.1]
  + 'max\_depth': [3, 5]

The top-performing parameters and corresponding metrics identified through this process are presented in the sections below.

### **7.1. Experiment 1: Binary Classification Models**

**Objective:** The primary experiment is designed to create the capability to properly categorize music in one of two different levels of popularity — either 0 or 1. This reflects the principal goals of the research, which are to identify and separate the factors of "hits" vs. "non-hits."

**Methodology:** Three models of classification were trained and tested against the processed database:

* Random Forest Classifier
* Support Vector Classifier (SVC)
* XGBoost Classifier

The dataset was split into training and testing sets to evaluate model performance on unseen data.

**Performance Evaluation:** The models' performance was assessed based on the evaluation methods described in Section 6.2 (Precision, Recall, and F1-Score). Each model's results were placed in a classification\_df (see Project Notebook for details) to assess their ability to perform the classification tasks. The F1-Score will be used to determine the most effective model since it is the most representative of both Precision and Recall.

**Preliminary Findings:** It was observed that the [e.g., Random Forest Classifier] had the most consistent performance among all three models and achieved the largest F1-Score. Therefore, it appears that the Random Forest Classifier was the most capable of appropriately identifying hit songs while limiting the number of misclassifications (false positives).

### **7.2. Experiment 2: Regression Models**

**Objective:** This experiment investigates another strategy — regression. Unlike the classification models previously examined, the goal of this experiment is to develop models that are capable of making predictions based on a continuous popularity rating (from 0-100) that existed in the original source data prior to binarization.

**Methodology:** Regression versions of the classification models were created to examine this possibility:

* Random Forest Regressor
* Support Vector Regressor (SVR)
* XGBoost Regressor

**Performance Evaluation:** For this type of problem, standard regression measures were used to determine the performance of these models as documented in the code used to generate the plots in the notebook:

* **R-squared (R2) Score:** Measures the proportion of the variance in the popularity score that is predictable from the features. A score closer to 1.0 indicates a better fit.
* **Root Mean Square Error (RMSE):** Measures the average magnitude of the errors in the model's predictions (in units of popularity). A lower RMSE is better.

**Preliminary Findings:** Although the regression models were able to provide some predictive power, the low R2 scores of the models suggest there may exist [e.g., considerable unaccounted variability]. This supported the researcher's hypothesis that predicting an exact "hit score" is very difficult. Therefore, the researchers decided that a classification approach (Experiment 1) was more suitable and effective to achieve the objectives of this study.

### **7.3. Experiment 3: Feature Importance & Explainability**

**Objective:** This last experiment focused on addressing the most critical portion of the project's goals developing "useful actionable information." (Vaughan, 2020). The objective of the project was to create an accurate model, but also to use that model to determine why certain characteristics contribute to a song's popularity.

**Methodology:** The best performing model from Experiment 1 (Random Forest Classifier) was chosen to be analyzed. The feature\_importances\_ attribute of the trained Random Forest model was obtained to determine the relative influence of each feature (e.g., danceability, energy, year of release) toward the classifier's prediction accuracy.

A screen shot of a computer code

AI-generated content may be incorrect.

Figure 22: Code for Feature Selection using Random Forest

**Performance Evaluation:** The "performance" of this experiment is measured by its explanatory power. The feature importances were ranked from highest to lowest.

**Preliminary Findings:** The analysis revealed the top-N key drivers of popularity as determined by the model. For example, features such as [e.g., danceability, loudness, and release\_year] were found to be highly influential, while [e.g., key or mode] had a less significant impact.

These findings are the direct inputs for the strategic recommendations and conclusions of this report, as they provide an empirical, data-driven answer to the core research question: "What makes a song popular?"

A screenshot of a computer

AI-generated content may be incorrect.

Figure 23: Random Forest Feature Importance Rankings

## **8. Cross-Experiment Summary**

The purpose of this Chapter is to summarize and compare the results of all the predictive models developed throughout this project. All the experiments were classified as either: binary Classification (determine if a song will be popular or not) or Regression (calculate a specific popularity score) and determine which one performs the best to achieve the Research Objective.

### **8.1. Binary Classification Top Results**

The classification experiments were run on both 80/20 and 70/30 data splits. The 80/20 split (test\_size=0.2) produced the most effective models.

**1. RandomForestClassifier**

**Best Parameters:** {'max\_depth': 20, 'min\_samples\_split': 5, 'n\_estimators': 150}

**Test Accuracy:** 89.75%

**Mean CV Score:** 87.89%

**2. GradientBoostingClassifier**

**Best Parameters:** {'learning\_rate': 0.1, 'max\_depth': 5, 'n\_estimators': 150}

**Test Accuracy:** 87.37%

**Mean CV Score:** 86.77%

**8.2. Regressor Top Results**

As with classification, the regression models performed best on the 80/20 split. While the models were successfully tuned, their overall R2 scores were very low (all below 0.19), indicating that the features are not effective for predicting a precise numerical popularity score.

**1. XGBoostRegressor**

**Best Parameters:** {'learning\_rate': 0.1, 'max\_depth': 7, 'n\_estimators': 100}

**Test R2 Score:** 0.1866

**Test RMSE:** 18.19

**Mean CV R2 Score:** 0.1604

**2. GradientBoostingRegressor**

**Best Parameters:** {'learning\_rate': 0.05, 'max\_depth': 5, 'n\_estimators': 150}

**Test R2 Score:** 0.1757

**Test RMSE:** 18.31

**Mean CV R2 Score:** 0.1623

## **9. Model Evaluation**

This final chapter evaluates the results of all experiments to select the best-performing model. It discusses the practical meaning of the model's findings, identifies the project's challenges, and provides a final conclusion for the study.

### **9.1. Best Models Across All Experiments**

The experiments were divided into two primary strategies: binary classification (predicting if a song's popularity is high or low) and regression (predicting the exact popularity score). Both strategies were tested using 80/20 and 70/30 data splits. The best algorithm from each category is summarized below.

### **9.1.1 Best Classifier Algorithm**

The **RandomForestClassifier** emerged as the top-performing model for classification.

* Using an 80/20 split (test\_size=0.2), this model achieved the highest **Test Accuracy at 89.75%**.
* It also demonstrated strong generalization with the highest **Mean Cross-Validation (CV) Score of 87.89%**.
* This performance was superior to both the GradientBoostingClassifier (Test Acc: 87.37%) and the SVC (Test Acc: 83.33%) under the same split.

A diagram of a model

AI-generated content may be incorrect.

Figure 24: Classification Model Accuracy Comparison

### **9.1.2 Best Regressor Algorithm**

For the regression task, the **XGBoostRegressor** provided the best results among the models tested.

* On the 80/20 split, it yielded the highest **Test R2 Score of 0.1866** and the lowest **Test RMSE of 18.19**.
* While it outperformed the SVR and GradientBoostingRegressor, it is critical to note that all regression models showed very low R2 scores (all below 0.19). This indicates that the features have minimal power in predicting an exact numerical popularity score, reinforcing the decision to focus on the more effective classification approach.

A diagram of a graph

AI-generated content may be incorrect.

Figure 25: Regression Model R2 Score Comparison

A graph showing a line

AI-generated content may be incorrect.

Figure 26: Regression Model RMSE Comparison

### **9.2. Recommendation**

Random Forest Classifier is the suggested last model in this research, due to the cross-evaluation among experiments.

Suggested model selection is based on two main elements:

1. **Predictive Performance:** Demonstrated highest performance in determining the target value..
2. **Interpretability:** Due to being an ensemble of decision trees, RFM is not a "black box"(Diakopoulos, 2014). RFM has a native attribute called feature\_importances\_. This is of greater strategic value than a slight increase in predictive accuracy for a model which is less transparent, as it enables transition from prediction to explanation..

### **9.3. Practical Recommendations**

Deployment of this project will be the conversion of model results into practical applications for artists, thereby achieving the goal of Strategic Thinking (Johnson and Scholes, 1993). Feature importance rankings derived from the recommended Random Forest Model serve as a data-driven reference point for music production and release strategies.

* **Insight 1: Focus on Contemporary Production (release\_year)**
  + **Finding:** The release\_year was a highly significant predictor.
  + **Recommendation:** The fact that listeners' preferences do not remain static; instead, they evolve. That older generations' production style is associated with lower popularity today. Therefore, artists must confirm that their production, mixing, and mastering are consistent with modern sonic standards.
* **Insight 2: Master for Perceived Energy (loudness & energy)**
  + **Finding:** The loudness and energy features were top drivers of popularity.
  + **Recommendation:** In relation to "loudness wars" and psychoacoustics. Songs that are perceived as having more energy (which is heavily related to their master-loudness levels) have a higher probability of being classified as "popular." Artists should spend money on quality mastering to make sure their songs compete with respect to perceived volume and energy on streaming platforms..
* **Insight 3: Engineer for Movement (danceability)**
  + **Finding:** Danceability consistently ranked as a key feature.
  + **Recommendation:** This does not imply all artists need to produce "dance" tracks. What it suggests is that songs that possess a strong, clear beat and groove – characteristics that cause them to receive a high danceability rating – have a statistical advantage. Therefore, producers should focus on creating clear and engaging rhythmic elements.

### **9.4. Challenges**

Even though the project produced favorable results, there are several challenges and limits that must be recognized:

* **Data Opacity:** The major problem was the static nature of the Kaggle dataset. The "high" and "low" popularity label criteria are unknown. The "unknown black box" label produces a considerable, unmeasurable restriction(Diakopoulos, 2014).
* **Missing External Variables:** The dataset is limited to audio variables and cannot represent the most influential determinants of success: marketing expenditures, playlists, social media virality (TikTok, etc.), artist reputation, and brand sponsorships.
* **Correlation vs. Causation:** This study reinforces a well-known weakness of data analysis. While the model indicates loudness correlates with popularity, it doesn't indicate it generates it. The insights obtained by this study are a representation of the statistical trend, not a promise of success.

### **9.5. Future Work**

In order to expand this research and to address the issues mentioned above, the following actions are suggested for future research:

* **Move to the Official Spotify API:** Use the official Spotify API: The next step is to replace the static dataset with the real-time data from the Spotify API. Using the true 0-100 popularity rating and the actual time data, this will enable the ability to obtain the real time data.
* **Integrate External Data Sources:** A more sophisticated model would ingest the data from the other APIs (Twitter, TikTok, etc.) to try to measure the influence of social media trends and marketing on the success of a song.
* **Apply Deeper Explainability (SHAP/LIME):** As mentioned in Chapter 4, using SHAP on the final model would provide the ability to perform instance-level explanations, thus answering the question "What makes songs popular?" and why the model believed that particular song would be popular
* **Analyze Lyrical Content:** The current model is "deaf" to the lyrics of the song. A future project may utilize Natural Language Processing (NLP) to analyze the content of lyrics (sentiment analysis, topic modeling, etc.) and include these variables in the model.

## **Use of Artificial Intelligence:**

Microsoft Copilot being integrated into Visual Studio Code greatly aided our team throughout the development phase; particularly for debugging since Copilot offered fixes and suggestions to help quickly resolve bugs and Copilot also made it much easier to generate useful comments and auto complete them. In addition to coding, the use of GPT-4.1 and Perplexity AI were used for many other aspects of the project including, but not limited to brainstorming and defining the initial architecture of the project.

Prompt example: “Giving this visualization that I created, what can I do to improve the design and image for better reading”; “How can I find a format for print the content more clearly, I want to organize the print content on the screen for better understanding”; “Help me to organize this markdown text”

## **References List (Harvard Style)**

1 - Davenport, T.H. and Harris, J.G. (2007). *Competing on Analytics*. Harvard Business Press.

2 - Diakopoulos, N. (2014). ‘Algorithmic Accountability: the investigation of Black Boxes’, Tow Center for Digital Journalism.

3 - FrancescaLazzeri (2022). ‘Deep learning vs. machine learning - Azure Machine Learning’.

4 - Gupta, B.B. and Mamta (2023). *Big Data Management And Analytics*. World Scientific.

5 - Indeed Career Guide (n.d.). ‘How To Calculate Statistical Significance (Plus What It Is and Why It’s Important)’.

6 - Johnson G. and Scholes K. (1993). *Exploring Corporate Strategy*. Prentice Hall.

7 - KATERI, ALAN. and Kateri, M. (2024). *Foundations of Statistics for Data Scientists*. CRC Press.

8 - Porter, M. E. (2008). ‘The Five Forces that Shape Strategy’, *Harvard Business Review*.

9 - Reis, J. and Housley, M. (2022). *Fundamentals of Data Engineering*. ‘O’Reilly Media, Inc.’.

10 - Vaughan, D. (2020). *Analytical Skills for AI and Data Science : Building Skills for an AI-Driven Enterprise*. Sebastopol: O’Reilly Media, Incorporated.