# CS 210 Data Analysis Project

# Spotify Music Listening Habits Analysis

Name and Surname:
**Elif Elmas**

Date:
**19.01.2024**

Course Name/Number:
**Introduction to Data Science CS 210**

Instructor's Name:
**Onur Varol**

Table of Contents

# 1. Executive Summary

In this study, "Spotify Music Listening Habits Analysis," listening preferences for different periods of the year are examined, with a special emphasis on the distinctions between the summer and school seasons. To identify seasonal trends in music preferences, the study analyzes my own Streaming History Spotify data, including genres and timestamps.

The methodology involved a detailed data collection process, enriching Spotify streaming history with genre information. In the feature engineering part, the data was categorized into 'school period' and 'summer', which prepared the model for training. Based on these periods, a Random Forest Classifier model was created to categorize different musical genres. The Mann-Whitney U test was used to test the hypothesis, and the results showed that there are notable differences in genre preferences between the summer and school periods.

The study's findings show an evident shift in my music preferences, indicating a tendency for deeper, introspective choices during the school period and vibrant, energetic selections in summer. The research resulted in the development of a predictive model designed to predict future music listening trends based on identified patterns. In addition to expanding my knowledge of my seasonal music preferences, this analysis develops a methodological framework that could be used in future research on digital consumer behavior and music psychology in larger studies.

# 2. Introduction

## 2.1. Background

Through platforms like Spotify, individuals' listening trends could be studied to reveal interesting insights into both popular culture and individual tastes. Using Spotify's large dataset, this study analyzes the potential impacts of seasonal shifts and the structure of the school period on listening habits and preferences. Offering useful data on user involvement and behavior, this study is relevant to a wide range of fields, including music psychology and digital music business marketing techniques.

## 2.2. Project Objective

The purpose of the study was to examine potential shifts in preferences for different music genres at different periods of the year, especially between the school period and the summer period. The project's goal was to find significant trends and preferences that occur in connection with these two periods' distinct habits by analyzing streaming data.

### 2.3. Hypothesis

The investigation was guided by the hypothesis (H0) "I hypothesize that there is a significant variation in my music listening habits, specifically in genre preferences, between the school period (excluding June, July, and August) and the summer months.", which posits a significant variation exists in music genre preferences between the school period and the summer period. The hypothesis was tested against the null hypothesis (H0) "There is no significant difference in genre preferences between the school period and the summer months.", which suggests there is no significant difference in music preferences between these two periods. This provided the basis for analyzing the seasonal shifts in music habits of listening using an analytical method.

## 3. Methodology

### 3.1. Data Collection

The Spotify data, which includes an exhaustive list of timestamps and genres, was meticulously collected from my streaming history. A high degree of accuracy and detail was ensured using structured JSON files and Spotify's API for obtaining the dataset. This data, which captured the nuances of my listening habits over nearly one year, functioned as the frame of the project.

### 3.2. Data Preprocessing

Extensive preprocessing of the data ensured it was ready for analysis. This was converting timestamps from 'endTime' into the standard DateTime format so that listening patterns could be evaluated in chronological order. Strict procedures were followed to ensure the accuracy of the data, and any missing values were filled in and duplicate entries were removed. The critical 'genres' variable was classified, making it more suitable for the upcoming analytical procedures.

### 3.3. Exploratory Data Analysis (EDA)

The EDA phase utilized techniques to find initial trends in the dataset by utilizing the features of Python's analytical libraries. A systematic analysis of genre frequencies over months was given by applying a pivot table, and temporal trends in musical preference were visually displayed by heatmaps. These exploratory measures had an essential function in shaping the direction of the following comprehensive analysis.

### 3.4. Feature Engineering

Creating new features was essential to matching the dataset to the study's hypothesis. The seasonal context of each listening event was assigned by applying the 'Season' from the 'endTime' data. Based on the month of each track, the 'Period' feature is separated between the 'School Period' and 'Summer'. To properly divide the data and enable a detailed examination of seasonal listening trends, these created features were needed.

## 4. Results

### 4.1. Model Selection and Training

A RandomForestClassifier model was chosen for handling the study problem due to its accuracy while processing categorical data and its ability to offer useful insights into genre preferences. The decision was taken considering the data's characteristics, which were mostly binary classifications into "School Period" and "Summer," along with categories genre classifications. The dataset that was used to train the model had undergone extensive preprocessing, ensuring the accuracy and quality of the input data. To train the model to recognize patterns in music preferences, the data was divided into training and testing sets. The testing set was reserved to assess the model's predicting abilities.

### 4.2. Hypothesis Testing

The trained RandomForestClassifier and the Mann-Whitney U test were both used to test the hypothesis. The model was used to determine the probability that a given genre will be more popular over the summer period or during the school period. In addition, a statistical evaluation of the variations in listening to music duration between the two periods was given by the Mann-Whitney U test. Determining whether there was a significant difference in music genre preferences between the school period and summer depended heavily on the data from both methodologies.

### 4.3. Model Evaluation and Refinement

A confusion matrix, accuracy scores, and classification reports were all a few metrics used to carefully evaluate the RandomForestClassifier's performance. These metrics gave a comprehensive insight into how well the algorithm classified genres within the given periods. The model was also subjected to a hyperparameter tuning procedure with GridSearchCV to improve its settings for better performance. In the end, this process of refinement helped to provide a more accurate representation and comprehension of listening habits.

# 5. Reflection and Future Directions

## 5.1. Interpretation of Results

The analysis of the study clearly shows there are differences in my music taste between the school period and the summer period. This claim was supported analytically by the RandomForestClassifier model and Mann-Whitney U test. The findings show a distinct pattern; during the school period, I prefer more reflective and potentially inspirational genres like rap, but during the summer period, I prefer a more vivid and varied variety, such as "ambient lo-fi," "urbano latino," and "reggaeton." This result shows that my music tastes are highly influenced by seasonal and environmental effects, which may be a reflection of various conditions and mental states.

## 5.2. Limitations

Despite being comprehensive, the research concurs with several shortcomings. The main limitation is the size of the data, which comes from the Spotify history of only one person which is me. This limitation raises concerns about the findings' ability for generalization. Moreover, the majority of the study is focused on genre preferences, ignoring other potentially important factors like mood, or external social events.

## 5.3. Reflection on Predictive Analysis

By proving the model's potential value in forecasting preferences, the experiment expanded its reach by projecting future music listening patterns. Personalized music suggestions become possible with the addition of this predictive analysis, which gives the study a forward-looking aspect. However, the accuracy of the data and the adaptability of the model are fundamentally related to how accurate these forecasts are.

## 5.4. Potential Improvements and Future Work

A more diversified dataset with multiple users might significantly improve future versions of this study. This extension would improve the reliability of the findings. Furthermore, adding more dynamic factors to the study, such as cultural trends or changes in mood over time, might improve it. Deeper insights into the nuances of the evolution of musical taste may be obtained by advancing the model by adding advanced algorithms or artificial intelligence approaches. These developments could lead to more personalized music recommendation algorithms on digital platforms.

## 6. Conclusion

The "Spotify Music Listening Habits Analysis" study provides an extensive examination of variations in music preferences over the year, with a particular focus on the summer vs the school period. The study is based on a comprehensive review of my individual Spotify Streaming History Data. The study has discovered significant shifts in genre preferences during these periods. The primary result of the research is the significant impact of seasonal variations and the academic calendar on my genre preferences. A distinction in my music taste is that I prefer more intense tracks throughout the school period and more positive, energizing tracks over the summer period. The use of the RandomForestClassifier and the Mann-Whitney U test, which provided statistical evidence for the main hypothesis, greatly enhanced the study's accuracy. While the predictive study produced some interesting results, it is currently limited by the quantity of the dataset and the model's capabilities, offering potential for further development and improvement. The digital music market continues to evolve and this attempt into the analytics of listening habits lays the basis for more advanced and directed methods.

## 7. References

Spotify AB. (n.d.). Spotify for Developers. from https://developer.spotify.com/documentation/web-api/