Data Introduction

The dataset used in this analysis is a music streaming dataset that comprises 12.5 GB of data, 26 million records, and 18 features. The dataset consists of various points of user interaction like session lengths, song preferences, demographics, and technical interactions like the type of device and user location.

* Features: The features present in the dataset are:
* User attributes: User ID, gender, user agent, and location.
* Session-based features: Session length, number of sessions, interaction time.
* Page-level attributes: The pages that are being viewed, i.e., certain cancellation-specific pages (churn prediction needed).
* Temporal features: Time stamp data regarding when events took place.

Churn Definition: For our project, **customer churn happens when a user reaches the "Cancellation Confirmation" page, which means they have cancelled their subscription.**

Problem Formulation

* Objective: The overall objective of this project is to **predict customer churn for a music streaming company.** Through knowing which customers are most likely to unsubscribe, the company can do something to keep these customers and minimize overall churn.
* Business Impact: Churn reduction is imperative to maximize customer retention and maximize the lifetime value of subscribers. Keeping users in the music streaming competitive market can directly influence profitability and long-term viability.
* Target Variable: The target variable of this study is churn, a binary classification problem in which we aim to predict if a user will churn (1) or not (0).

Descriptive Analysis

* The dataset has 26,259,199 instances and 18 attributes.
* There are 22,278 unique users in the data.

Missing Values:

* Certain features contain missing values, which are addressed during preprocessing by imputing them with suitable values (e.g., with the mean for numeric features) or omitting them where required.

Dropout Analysis:

* 5003 churn events were discovered from the dataset, a churn rate of 22.46% of the users.
* A churn event is recorded when the user is on the "Cancellation Confirmation" page.

- Feature Distributions: - statistics for numeric features such as session length, number of sessions, and active length are given. These statistics give the first insight into user engagement and direct the remainder of the modelling.