DataWave Music Dataset Guide

Welcome to the **dataset guide** for the IoA Student Sprint: The Analytics Task Challenge. This document is here to help you understand what's in your dataset, how to explore it, and how to think about it critically. The dataset has been designed to resemble real-world data, meaning it's not perfect, but it's full of opportunities to uncover meaningful insights.

1. About the Dataset

You've been given a fictional dataset from **DataWave Music**, a global music streaming company. This dataset contains information about users, their subscription types, listening behaviour, and satisfaction ratings. Your task is to **analyse this data** and **find patterns** that can help the business **improve customer engagement** and **retention**.

In a real analytics role, data like this might come from user activity logs, survey feedback, or subscription databases. The aim is to give you a **realistic challenge** that reflects what analysts and data scientists encounter day-to-day.

2. Dataset Overview

The dataset contains roughly 700 rows and 12 columns. Each row represents an individual user. You'll find a mixture of numerical, categorical, and date-based fields.

Broadly, the dataset captures:

- User details (e.g. ID, country, age).
- Subscription and engagement behaviour (e.g. hours listened, plan type).
- Customer satisfaction and churn outcomes (e.g. survey scores, whether they cancelled).

3. Data Dictionary

Column Name	Description	Data Type
user_id	Unique identifier for each user (may contain duplicates or missing values).	Integer / Text

country	The user's country of residence.	Categorical (Text)
subscription_type	Type of subscription plan (Free, Premium, Family, Student, etc.).	Categorical (Text)
age	Age of the user, where available.	Numeric
hours_listened	Average weekly listening time.	Numeric
favourite_genre	The user's most listened-to genre.	Categorical (Text)
device	The primary device used to access the platform (e.g. Mobile, Desktop, Tablet).	Categorical (Text)
satisfaction_score	Customer satisfaction score from survey (1–10 scale).	Numeric
churn	Whether the user has cancelled their subscription (1/0, Yes/No).	Boolean / Text
date_joined	Date the user joined the platform.	Date
last_active	Most recent date of activity.	Date
region	Regional grouping of countries (e.g. Europe, Asia, Africa).	Categorical (Text)

4. Known Data Quality Issues

The dataset intentionally includes a few imperfections to give you the opportunity to practise real data-cleaning tasks. You may encounter:

• Inconsistent spellings or formatting in country names and subscription types.

• Missing or duplicated user IDs.

• Mixed date formats (e.g. DD/MM/YYYY and MM/DD/YYYY).

• Non-numeric text in numeric columns (e.g. 'ten' instead of 10).

• Outliers or unrealistic values in fields such as age or hours listened.

There may be other errors that need to be corrected…

5. Suggested Steps to Begin

1. Familiarise yourself with the dataset by exploring each column and checking for missing or inconsistent data.

2. Clean the data as needed, correct misspellings, handle missing values, and standardise categories.

3. Start exploring relationships between variables. For example:

- Does satisfaction differ by subscription type or region?
- Are there patterns between listening hours and churn?
- Which groups of users seem the most engaged or likely to leave?

4. Create visuals that help tell the story. Charts, tables, and dashboards are all welcome.

5. Summarise your findings in a way that a business leader would understand, focus on what the data means, not just what it shows.

6. Ethical and Communication Reminders

• Treat the dataset as professional data. Avoid making assumptions about individuals or using inappropriate language.

• Be transparent in your communication. If you make an assumption or use an estimation, mention it.

• Focus on clarity and impact. Employers value analysts who can explain data clearly and responsibly.

7. Common Pitfalls

• Assuming correlations imply causation… explore relationships carefully.

• Ignoring outliers without investigating their cause. Often they're the most insightful points.

• Overloading your presentation with visuals. Use them purposefully to support your key points.

• Forgetting to check consistency between variables such as churn, satisfaction, and activity dates.

8. FAQ

Q: Can I use Excel or Tableau instead of coding?

A: Yes. You can use whichever tool you're most confident in. You'll be assessed on clarity and interpretation, not tool complexity.

Q: What if I find an error in the data?

A: Note it in your presentation. Handling imperfect data thoughtfully is part of the challenge.

Q: Do I need to show my code or data-cleaning steps?

A: No, but you may include a brief overview or screenshots if it helps explain your approach.

Q: Should I submit the cleaned dataset?

A: No, only your presentation deck and recorded walkthrough are required.

9. Final Note

There's no single 'right' way to analyse this dataset. The goal is to show how you think, how you interpret information, and how you communicate findings. The best analyses usually balance logic, curiosity, and storytelling.

Good luck! Enjoy uncovering the story hidden within the DataWave Music data.

Rohan Whitehead
Data Training Specialist
Institute of Analytics