# DATA420 Assignment 2 - Progress Overview

Date: October 9, 2025

Repository: David-Ewing-82171165-DATA420-A2

Branch: main (commit: d2a7f0a)

**Current File:** 20251009A-A2-Processing.ipynb **Status:** Q2 Complete, Executed Successfully

# SECTION 1: DATA PROCESSING (20% of Assignment)

✓ Q1: Directory Tree & Statistics - **COMPLETE** 

### Q1(a): Create directory tree diagram

• ✓ Output: msd\_directory\_tree.png

#### Q1(b): Compute file statistics

• Files: 9 datasets analyzed

#### Q1(c): Display as formatted table

• Columns: 11 attributes per dataset

## ☑ Q2: Audio Feature Processing - **COMPLETE & EXECUTED**

#### Q2(a): Attribute Analysis

- Parse 9 datasets (3929 total columns)
- Collision detection (0 found)
- Type mapping (string/real)

#### Q2(b): Automatic Schema Generation

- ✓ Helper: create\_struct\_type\_from\_attributes()
- ✓ Generated 4 schemas (185 total fields)
- Types: StringType, DoubleType

#### Q2(c): Column Naming Discussion

- Advantages: self-documenting, traceable
- Disadvantages: too long (avg 18, max 108)
- 🗹 Conclusion: systematic renaming needed

#### Q2(d): Systematic Column Renaming

- ✓ Naming format: {AA}{NNN} (5 chars fixed)
- ✓ Helper: rename\_audio\_columns()
- 4 datasets renamed (180 features)
- **☑** 72% character reduction
- ✓ Mapping CSV: audio\_column\_name\_mapping.csv

### **■** Column Naming Convention

- A0 = Area-Of-moments (AO001–AO020)
- LP = LPc (LP001–LP020)
- SP = SPectral-all (SP001–SP016)
- TI = TImbral (TI001–TI124)
- MSD\_TRACKID preserved as join key

## ■ SECTION 2: AUDIO SIMILARITY (40% of Assignment)

CURRENT FOCUS: Starting Audio Similarity Analysis

**Z** Q1: Binary Classification Prep - **PENDING** 

### Q1(a): Load 4 renamed datasets

\( \sum \) Use: \( \text{renamed\_dfs} \) dict from Q2(d)

#### Q1(b): Create correlation heatmap

- ☐ Output: Figure 2 (correlation matrix)
- Identify highly correlated features

#### Q1(c): Remove correlated features

- ☐ Threshold: |correlation| > 0.9
- Z Create reduced feature set

#### Q1(d): Join datasets + popularity

- ☐ Create binary labels (popular/not)

## 🔀 Q2: Binary Classification Models - **PENDING**

#### Train 3 models:

- Random Forest
- 🖫 Gradient-Boosted Trees

#### **Evaluate with:**

- ■ ROC-AUC scores
- ■ Confusion matrices
- X Feature importance

### **Z** Q3: Multiclass Genre Classification - **PENDING**

#### Prepare genre labels:

- 🔀 Join with audio features
- ☐ Handle class imbalance

#### **Train multiclass models:**

- ■ OneVsRest strategy
- Z Per-genre metrics

# **I** SECTION 3: SONG RECOMMENDATIONS (40% of Assignment)

- ☑ Q1: ALS Collaborative Filtering PENDING
  - Matrix factorization for recommendations
- **Q**2: Content-Based Filtering **PENDING** 
  - Audio feature similarity recommendations
- **Q**3: Hybrid Approach **PENDING** 
  - Combine collaborative + content-based
- ☑ Q4: Evaluation & Comparison PENDING
  - Compare all recommendation strategies

## FILE HISTORY & COMMITS

- ✓ 20251008C-A2-Processing.ipynb
  - **Commit:** 36d216c
  - Content: Q1 + Q2(a) with outputs
- ✓ 20251008D-A2-Processing.ipynb
  - Commits: fc7dfe4 → d2a7f0a
  - Content: Added Q2(b), Q2(c), Q2(d)
  - Fix: Fixed cprint() errors

## 20251009A-A2-Processing.ipynb ← CURRENT WORKING FILE

- Status: All Q2 executed successfully
- Next: Ready for Audio Similarity section

## F KEY ARTIFACTS GENERATED

- 1. ✓ msd\_directory\_tree.png Visual directory structure (Q1a)
- 2. ✓ audio\_column\_name\_mapping.csv 185 rows: original → new column names (Q2d)
- 3. renamed\_dfs dictionary 4 DataFrames with renamed columns (Keys: 'AO', 'LP', 'SP', 'TI')
- 4. ✓ schemas dictionary 4 StructType schemas for CSV loading
- 5. ✓ all\_mappings dictionary Column name translation tables

# ## HELPER FUNCTIONS (Cell 8, Lines 253–1078)

### Main Q2 Functions:

- 1. create\_struct\_type\_from\_attributes() Maps attribute lists to Spark StructType schemas
- 2. rename\_audio\_columns(df, code, keep\_msd) Renames columns to {AA}{NNN} format
  - Returns: (renamed\_df, mapping\_dict)

### Plus 20+ additional helpers from Q1/Q2(a):

- hprint() formatted headers
- File size calculation functions
- Table formatting utilities

## **EXECUTION METRICS**

### Cells:

- Total Cells: 36
- Cells Executed: 36
- **Errors:** 0

#### Data Processing:

- Datasets Processed: 9 analyzed, 4 renamed
- **Total Columns:** 3929 original → 185 renamed
- Character Reduction: 72%

### Features by Dataset:

- AO: 20 | LP: 20 | SP: 16 | TI: 124
- Join Key: MSD\_TRACKID (preserved)

## **\*** NEXT IMMEDIATE STEPS

- Step 1: Download Mapping CSV
  - From Spark → local repository
  - Path: report/supplementary/

### Step 2: Commit 20251009A with Outputs

- git add + commit + push
- Include executed notebook

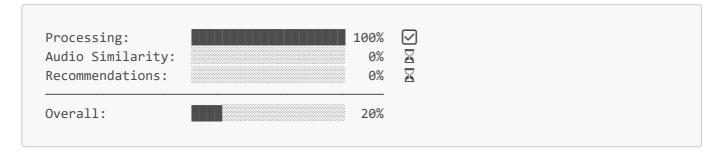
### 

- Load renamed\_dfs datasets
- Begin correlation analysis

### 

- Correlation heatmap visualization
- Identify features to remove

# **M** OVERALL PROGRESS



## □ LEGEND

- Completed Executed and verified
- **Z Pending** Not yet started
- \* Fixed Corrected errors

## 

Generated: October 9, 2025

**Repository:** github.com/david-ewing-nz/David-Ewing-82171165-DATA420-A2 **Status:** All helper functions organized in Cell 8 | Zero errors | Ready for execution