Collins Tse Jacky Ke Rebecca Bachtra Christy Yau

Thursday 18th December, 2025

Curvenote

```
import os
import pandas as pd
import numpy as np

def load_and_clean_spotify(path="data/spotify.csv"):
    # Load CSV and skip malformed rows
    df = pd.read_csv(path, on_bad_lines="skip", low_memory=False)

    # Drop unnamed index column
    df = df.loc[:, ~df.columns.str.contains("^Unnamed")]

    # Strip whitespace in all strings
    for col in df.select_dtypes(include=["object"]):
        df[col] = df[col].astype(str).str.strip()

    # Fix booleans
    if "explicit" in df.columns:
        bool_map = {"True": True, "False": False, "1": True, "0": False}
        df["explicit"] = df["explicit"].map(bool_map)

    # Convert numeric columns
    numeric_cols = [
        "popularity", "duration_ms", "danceability", "energy", "key",
        "loudness", "speechiness", "acousticness", "instrumentalness",
        "liveness", "valence", "tempo"
    ]
    for col in numeric_cols:
        if col in df.columns:
            df[col] = pd.to_numeric(df[col], errors="coerce")

    # Remove duplicate track IDs
    if "track_id" in df.columns:
        df = df.drop_duplicates(subset=["track_id"])

    # Add new engineered features
```

```
    if "duration_ms" in df.columns:
        df["duration_min"] = df["duration_ms"] / 60000

    if "year" in df.columns:
        df["decade"] = (df["year"] // 10) * 10

    # Clean track_genre text
    if "track_genre" in df.columns:
        df["track_genre"] = (
            df["track_genre"]
            .str.replace(",.*", "", regex=True)
            .str.lower()
            .str.strip()
        )

    # Remove rows with missing popularity (required for modeling)
    if "popularity" in df.columns:
        df = df.dropna(subset=["popularity"])

    return df

df_clean = load_and_clean_spotify("data/spotify.csv")
df_clean.head()
```

	track_id	artists	album_name	track_name	popularity	duration_ms	explicit	danceability	energy	key	...	mode	speechiness	acousticness	instrumentalness	liveness	valence	tempo	time_signature	track_genre	duration_min
0	5SuOikwiRyPMVoIQDJUgSV	Gen Hoshino	Comedy	Comedy	73	230666	0	0.676	0.4610	1	...	0	0.1430	0.0322	0.000001	0.3580	0.715	87.917	4	acoustic	3.844433
1	4qPNDBW1i3p13qLCt0Ki3A	Ben Woodward	Ghost (Acoustic)	Ghost - Acoustic	55	149610	0	0.420	0.1660	1	...	1	0.0763	0.9240	0.000006	0.1010	0.267	77.489	4	acoustic	2.493500
2	1iJBSr7s7jYXzM8EGcbK5b	Ingrid Michaelson;ZAYN	To Begin Again	To Begin Again	57	210826	0	0.438	0.3590	0	...	1	0.0557	0.2100	0.000000	0.1170	0.120	76.332	4	acoustic	3.513767
3	6lfxq3CG4xtTiEg7opyCyx	Kina Grannis	Crazy Rich Asians (Original Motion Picture Sou...	Can't Help Falling In Love	71	201933	0	0.266	0.0596	0	...	1	0.0363	0.9050	0.000071	0.1320	0.143	181.740	3	acoustic	3.365550
4	5vjLSffimiIP26QG5WcN2K	Chord Overstreet	Hold On	Hold On	82	198853	0	0.618	0.4430	2	...	1	0.0526	0.4690	0.000000	0.0829	0.167	119.949	4	acoustic	3.314217

5 rows × 21 columns

```
os.makedirs("data/processed", exist_ok=True)

outpath = "data/processed/spotify_clean.csv"
df_clean.to_csv(outpath, index=False)

print("Saved cleaned dataset to:", outpath)
```

```
Saved cleaned dataset to: data/processed/spotify_clean.csv
```