King Saud University
College of Computer and Information Sciences
Department of Information Technology

Spotify

IT 362 Course Project
Semester-1, 1447H

Student Name	Student ID
Dana Alsalami	443200945
Norah Aljayan	444200832
Aljohara Alsultan	444203635
Layan Alhowaimel	444200969
Alya Almuqren	444200610

Course Instructor: **Dr. Abeer Aldayel**

Logbook Entry – Data Processing & Cleaning

Date: 2025-09-19

Steps Taken:

Checked for missing values and duplicates.

Removed duplicates using artist ID.

Reason: Duplicate entries could bias summary statistics by giving extra weight to artists with repeated entries or duplicate records.

Converted genres from list to text (genres_str) for analysis.

Created derived columns:

genre_count = number of genres per artist

genre_category = “single-genre” or “multi-genre”

Reason: To standardize the dataset and prepare for visualization.

Tools: Pandas

Challenges: Handling nested list-type columns; solved by converting to string.

Code snippets:

```
print("Missing values per column:\n", df_artists.isnull().sum())
```

Drop duplicates based on artist ID

```
df_artists.drop_duplicates(subset='id', inplace=True)
```

Convert genres list to comma-separated string for readability

```
df_artists['genres_str'] = df_artists['genres'].apply(lambda x: ', '.join(x))
```

Confirm datatypes

```
print("\nData types after cleaning:\n", df_artists.dtypes)
```

Logbook Entry – Exploratory Data Analysis (EDA)

Date: 2025-09-20

Analyses Performed:

- Summary statistics (mean, median, std).
- Correlation heatmap (genre count vs popularity).
- Boxplot: Popularity by genre category: Multi-genre artists generally show **slightly higher median popularity** than single-genre artists.
- Scatter plot: Popularity vs number of genres: Popularity tends to remain high among artists with **2–5 genres**, but there’s no clear linear increase.

- Histogram: Distribution of popularity:The dataset is dominated by **moderately popular artists**, with fewer extremes.
- Multi-genre artists tend to have higher popularity.
- Popularity distribution is right-skewed — a few artists dominate.
- Weak positive correlation (≈ 0.12) between genre count and popularity.
  Tools: Pandas, Seaborn, Matplotlib.

Logbook Entry – Primary & Secondary Data

Date:2025-09-20

Despite using an API, we are treating the data we collected as primary since we actively selected which artists to analyze and manually pulled their data from their spotify API and queried the API to generate our own dataset rather than relying on someone else's dataset in full.

For secondary data, we used Kaggle to download a dataset called Spotify Songs and Artists Dataset | Audio Features (link: **https://www.kaggle.com/datasets/glowstudygram/spotify-songs-and-artists-dataset?resource=download)**