Final Project Report

Data Acquisition

The initial data acquisition approach involved attempting to scrape music-related websites using traditional HTML web scraping techniques. However, these attempts were consistently blocked due to website security restrictions such as anti-scraping measures and access limitations, making reliable data collection infeasible. To address this issue, the Discogs website was selected as an alternative data source due to its publicly available and well-documented REST API. After registering for developer access, a consumer key and secret were obtained, allowing authenticated requests to the Discogs database. Using this API, structured music release data was programmatically collected, normalized, and transformed into a tabular format suitable for analysis, database storage, and visualization.

Data Cleaning

The data cleaning process began with an examination of missing values across all columns in the dataset. Although no explicit missing values (`NaN`) were detected, further inspection revealed the presence of empty strings in several categorical fields such as country, genre, style, and label. These empty strings were treated as incomplete data, and rows containing such values were removed to ensure consistency and reliability in subsequent analyses. Following this, duplicate records were assessed based on attributes like year, country, and format. While some similarities were observed across these fields, the records were retained because they represent distinct music releases and remain meaningful for trend analysis, aggregation, and country-level comparisons. This approach ensured the dataset remained comprehensive while maintaining data quality.

Data Analysis

The data analysis process began with an initial exploration of all available fields in the dataset, starting with identifying the different types of music genres. During this step, it became clear that the genre column was not stored as individual values, but instead contained comma-separated strings representing multiple genres per release. To properly analyze genre distribution, the column was transformed by

splitting these strings into individual genre values and extracting unique entries using a set-based approach. This allowed for a more accurate representation of genre diversity within the dataset.

A similar issue was identified in the format column, which also contained multiple formats stored as comma-separated values. The same data-splitting and normalization technique was applied to ensure each format could be analyzed independently. This approach was extended across several fields in the dataset, including style and label, enabling a more granular and meaningful analysis. By normalizing these columns, relationships between genres, styles, formats, labels, and countries could be examined more effectively.

With the cleaned and structured data, deeper analysis focused on the connections between genres and styles, as well as how these relationships vary across different record labels and countries. This revealed patterns indicating that certain genres and styles are strongly associated with specific regions and labels, suggesting localized music ecosystems and genre specialization. To further explore and communicate these relationships, the cleaned dataset was exported to an Excel (.xlxs) file and imported into Tableau. This enabled the creation of advanced visualizations that highlight genre–style dominance, label concentration, and regional music trends, providing a clear visual narrative of how global music production differs across countries and markets.

Database Storage

For the database storage component of this project, MySQL was selected as the primary data management system due to its reliability, structured schema design, and strong support for analytical querying. After completing the data cleaning and transformation steps in Jupyter Notebook, the finalized dataset was programmatically transferred into a MySQL database named Dsci105, where it was stored in a table titled discogs_music. This process was carried out using Python scripts with MySQL connectivity, ensuring that the data was accurately and consistently persisted.

To support ongoing analysis, additional scripts were implemented to retrieve the stored data back into the Jupyter environment as needed. This workflow eliminated

the need for repeated data acquisition and cleaning, enabling efficient exploration and analysis of the dataset. Once stored, SQL queries and pandas operations were used to examine the overall volume of music releases and to further analyze attributes such as genres, styles, formats, labels, years, and countries. This database-backed approach provided a stable foundation for systematic analysis and seamless integration with subsequent visualization tasks in Tableau.

Visualization with Tableau

The Tableau visualizations were created using a dataset of music releases collected from the Discogs API and stored in a structured SQL table. The dataset contains detailed information on release titles, years, countries, genres, styles, formats, and record labels. After cleaning and preparing the data, it was imported into Tableau, where multiple visualization techniques—including bar charts, stacked bar charts, heat maps, bubble charts, pie charts, and a geographic map—were used to explore patterns and trends. Interactive filters for country, genre, format, style, and label were applied across all dashboards, allowing users to dynamically explore the data from multiple perspectives.

Through these visualizations, several clear patterns emerge. Rock and Electronic music appear as the most dominant genres, with Rock showing the highest overall release count. Further breakdown at the style level reveals strong representation of subgenres such as Punk, Black Metal, and Hard Rock, highlighting the presence of active niche music communities. Format-based charts indicate a strong dominance of physical formats, particularly Vinyl LPs and CDs, suggesting continued interest in tangible music media. Country-level comparisons show that the United States, Germany, and the United Kingdom contribute the largest share of releases, while the global map demonstrates that genre popularity varies significantly across regions rather than following a single global trend.

Taken together, these patterns suggest that music production and distribution remain highly region-specific and influenced by cultural preferences and local markets. The continued prominence of vinyl and CD formats implies that physical releases still play an important role, especially within certain genres and countries. The dashboards enhance analytical depth by enabling users to filter and compare

countries, genres, and formats interactively. One dashboard provides a global overview of music releases by visualizing dominant genres, popular formats, and leading styles across regions, while the second dashboard focuses on country-specific analysis, examining genre distribution, record label presence, and national contributions. Overall, the Tableau dashboards offer a comprehensive and flexible framework for understanding global music release trends and demonstrate how visual analytics can reveal meaningful insights from complex datasets.