King Saud University

College of Computer and Information Sciences

Department of Information Technology

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**

Table of Contents

Table of Figures

Introduction

The music industry has witnessed increasing fluidity in genre boundaries, with many artists exploring multiple styles rather than remaining confined to a single category. This raises an important question about the relationship between artistic versatility and success: **Do artists who span multiple genres have higher popularity compared to those who focus on a single genre?**

To address this research question, we employ data retrieved from Spotify Charts and enriched with the Spotify API. First, we collected weekly regional chart CSVs from multiple countries, each containing the top 200 tracks. From these charts, we extracted unique artist names. Using a custom script, we then queried the Spotify API to gather additional metadata for each artist, including their Spotify ID, associated genres, and popularity scores (0–100 scale). This integrated dataset forms the basis for analyzing whether artists spanning multiple genres tend to achieve higher popularity among listeners.

The study will evaluate popularity trends across multi-genre versus single-genre artists, and highlight implications for the modern music industry. The overall aim is not only to assess whether versatility contributes to greater popularity, but also to offer insights into how genre diversity and digital streaming platforms shape contemporary musical success.

Data sources

For our project on artist popularity and genre diversity, we relied primarily on two data sources: **Spotify Charts** and **Spotify Web API.**

Spotify Charts (CSV files)

Spotify Charts are weekly and daily charts published by Spotify, listing the top ten tracks in various countries and globally. We downloaded the weekly top tracks of the following countries: Global, US, Saudi Arabia (SA), Belarus (BY), Italy (IT), Brazil (BR), India (IN), United Kingdom (UK), Taiwan (TW), Switzerland (CH), South Korea (KR), Japan (JP), Venezuela (VE), Uruguay (UY), Turkey (TR), Thailand (TH), Australia (AU), United Arab Emirates (AE), Hong Kong (HK), Egypt (EG).

Data collected: Artist names, track names, chart rank, streams, weeks on chart, peak rank.
Number of observations: After consolidating all downloaded CSV files, there are approximately 1785 unique artists.

Features & types:

Track_name (string): Name of track

artist_names (string): Name of the artist)

rank(integer): Current rank of the chart (integer)

streams(integer): Number of streams of the chart

weeks_on_chart(integer): Number of weeks the track has been on the chart

peak_rank(integer): Highest rank achieved on the chart

previous_rank(integer): the tracks position in the previous week

source (string): chart source (country/region) The only relevant feature to our hypothesis is the artist_names feature

Spotify API

We used the Spotipy Python library with a developer Spotify account to retrieve detailed artist information.

Data collected: Artist ID, genres, and popularity metric

Features & types: id (String): Spotify unique identifier for the artist

name (string): artist name

genres (list): list of genres associated with the artist

popularity(integer): Spotify popularity score (0-100) calculated based on recent streams and engagement

Representation Bias: Each chart contains 200 entries, but the underlying listener populations differ across regions. Large markets like the US, UK, and Global charts reflect millions of streams, while smaller markets like Uruguay or Thailand reflect fewer streams.The dataset includes only **unique artists**, so each artist appears once even if present in multiple charts. This reduces the overrepresentation of globally popular artists. However, artists from smaller markets are still less likely to appear, and genres or languages favored in large markets may dominate.

Measurement Bias: Spotify popularity is influenced by recent streams and engagement. New releases may temporarily inflate an artist's popularity, while older or niche artists may be underrepresented.

Data Limitations: Artists with multiple pseudonyms or collaborations may appear under different names. Artists without genres listed in the Spotify API are excluded, biasing the dataset toward mainstream or well-categorized artists.

Objectives

1. **To investigate the relationship between genre diversity and popularity** by comparing artists who span multiple genres with those limited to a single genre.
2. **To collect and analyze data from the Spotify API**, including artist names, IDs, genres, and popularity scores, in order to build a reliable dataset for evaluation.
3. **To identify potential trends or correlations** between cross-genre versatility and higher popularity metrics.
4. **To examine possible biases or limitations** in the dataset, such as unequal representation of artists or genre classifications.
5. **To provide insights into the role of genre diversity** in shaping artist success and its implications for the modern music industry

Method

Data Collection

Weekly Spotify chart data from 20 regions, including Global, US, UK, Saudi Arabia, Japan, Brazil, and others, were collected by downloading CSV files containing track names, artist names, chart ranks, streams, weeks on chart, and peak positions. The CSV files were programmatically imported and merged into a single dataset using Python. Duplicate artist entries were removed to ensure that each artist appeared only once across all charts. From this consolidated dataset, all unique artist names were extracted. Using the Spotipy Python library with Spotify Web API credentials, detailed artist metadata was retrieved for each unique artist,

including Spotify ID, genres, and popularity scores. Artists without genre information or duplicate IDs were excluded to maintain data consistency. Error handling was implemented to manage connection errors and API exceptions during data retrieval. The resulting dataset was stored in a Pandas Data Frame containing only artists with non-empty genres and valid popularity metrics.

Data Preprocessing

Preprocessing consisted of merging multiple CSV files, filtering out duplicate artists, and removing any artists with missing genre information. This ensured that the dataset was consistent and ready for analysis. All files were stored locally to allow reproducibility without repeated API requests.

Data Analysis and Visualization by Objective

Objective 1: *Investigate the relationship between genre diversity and popularity.*

- We created a new feature genre_count to indicate how many genres each artist spans. Artists were categorized as **single-genre** or **multi-genre**. Popularity distributions were compared using **boxplots**, and the number of artists per category was visualized using **countplots**.

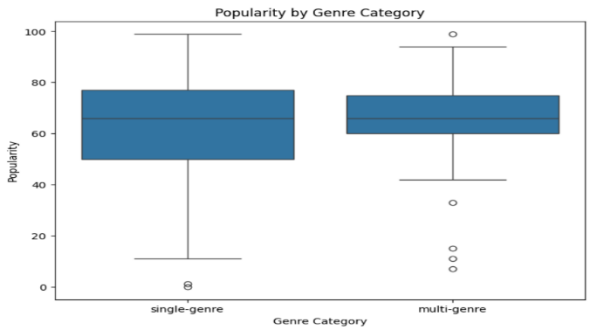

Figure 1:Boxplot

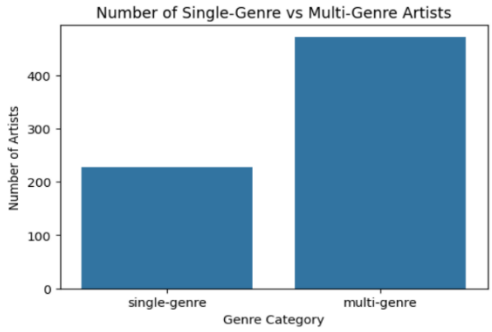

Figure 2:Countplot

Objective 2: *Collect and analyze data from the Spotify API, including artist names, IDs, genres, and popularity scores.*

- CSV files from 20 regions were merged, and unique artist names were queried from the Spotify API using Spotipy to retrieve Spotify IDs, genres, and popularity scores. Artists

without genres or duplicates were excluded. The final dataset was stored in a Pandas DataFrame for analysis.

Objective 3: *Identify potential trends or correlations between cross-genre versatility and popularity metrics.*

- Correlation between genre_count and popularity was examined. Trends were visualized using **scatter plots** and **regression plots** to assess whether spanning more genres is associated with higher popularity.

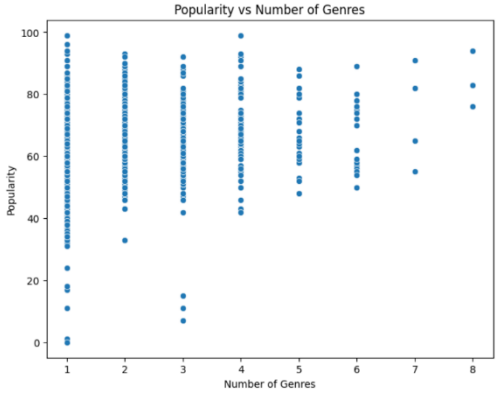

Figure 3:Scatter plot

Objective 4: *Examine potential biases or limitations in the dataset.*

- The dataset includes only unique artists across all charts, which reduces the overrepresentation of globally popular artists but also means that artists from smaller regions are less likely to appear overall. Since each chart contains 200 entries but represents very different listener populations, genres and languages from larger markets (such as the US or UK) may dominate the dataset compared to smaller markets like Uruguay or Thailand. Spotify's popularity score is also shaped by recent streams, so new releases may temporarily inflate popularity while older or niche artists are underrepresented. Finally, artists without listed genres in the Spotify API were excluded, which introduces a bias toward mainstream or well-categorized artists, and artists using

pseudonyms or frequent collaborations may appear multiple times under different names. Visualizations such as **countplots** of genre categories help illustrate potential representation biases.

Objective 5: *Provide insights into the role of genre diversity in shaping artist success.*

- Descriptive statistics, histograms of popularity, boxplots, scatter plots, and regression plots were used to explore and compare popularity across single-genre and multi-genre artists. These analyses reveal patterns linking genre diversity to popularity and provide insights into its impact on modern music industry trends.

Tools and Libraries

Python libraries used in the project include:

- **spotipy**: Retrieve artist metadata from Spotify API
- **pandas**: Data manipulation and DataFrame creation
- **os** and **glob**: File handling and importing multiple CSVs
- **requests**: Downloading CSV files
- **time**: Managing pauses between API requests
- **spotipy.exceptions** and **requests.exceptions**: Error handling during API calls
- **matplotlib** and **seaborn**: Data visualization for descriptive and comparative analysis

Challenges

1. API Limitations and Errors

Collecting data for hundreds of artists required many requests, which often hit rate limits or caused timeouts.

2. Incomplete Genre Information

Some artists did not have genre data listed in the Spotify API, leading to their exclusion and reducing the dataset's diversity.

3. Representation Bias

Larger markets (e.g., US, UK, Global) dominated the charts, while smaller markets were underrepresented, potentially biasing results toward globally popular genres.

4. Limited Data for Emerging Artists

Less popular or newer artists often have incomplete data or fewer genre tags, which can bias comparisons between artists with many genres and those with few.

5. Duplicate/alias issues

Artists with multiple pseudonyms, collaborations, or spelling variations may appear under different IDs, fragmenting the dataset.

Primary Data :

Contextual Information

The Primary dataset was compiled using two main sources:

1. **Spotify Weekly Regional Charts** (20 regional CSV files) obtained from GitHub.
2. **Spotify Web API** accessed via the Spotipy library.

Each artist entry includes their **Spotify ID**, **popularity score**, and **associated genres**. After cleaning and merging, a single dataset (df_artists) was created, containing artist-level metadata suitable for exploratory analysis.

EDA Results

Data Cleaning & Preparation

- Removed duplicate entries using artist ID.
- Handled missing values (artists with missing genres were excluded).
- Converted list-type genre fields into strings (genres_str).
- Created two derived features:
    - genre_count: number of genres per artist.
    - genre_category: "single-genre" or "multi-genre"

Descriptive Statistics by Genre Category

Max	75%	50% (Median)	25%	Min	Std	Mean Popularity	Count	Genre Category
95.0	71.0	61.0	55.0	5	12.75	62.04	475	Multi-genre
95.0	72.0	61.0	45.0	0	18.75	58.01	233	Single-genre

Interpretation:

This table shows that multi-genre artists tend to be slightly more popular than single-genre artists, with a higher mean popularity (62.04 vs 58.01) and lower variability (std = 12.75 vs 18.75), meaning their popularity is more consistent.
Both groups share a similar median popularity (61.0) and maximum score (95.0), but single-genre artists show a wider range (0–95) compared to multi-genre (5–95), indicating a greater spread in popularity levels among single-genre artists.

Visual Findings

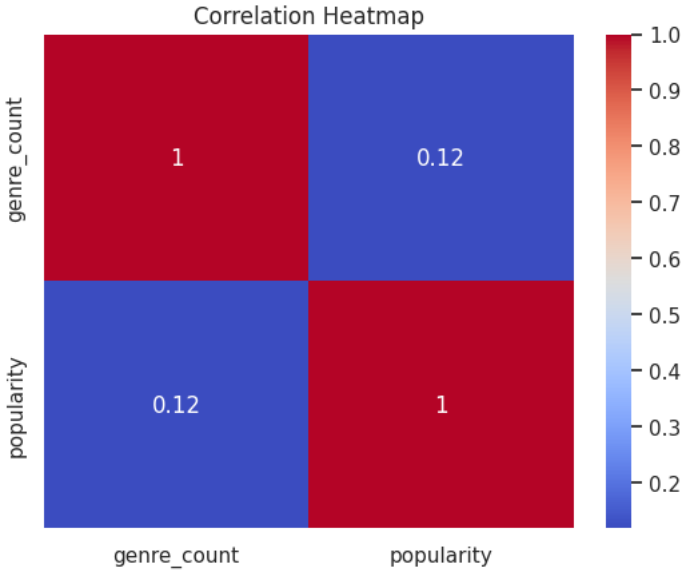

Figure 4: Primary Data Heatmap

This heatmap shows the correlation between "genre_count" and "popularity." The correlation coefficient is **0.12**. This very low value confirms a very weak positive linear relationship between the number of genres an item has and its overall popularity.

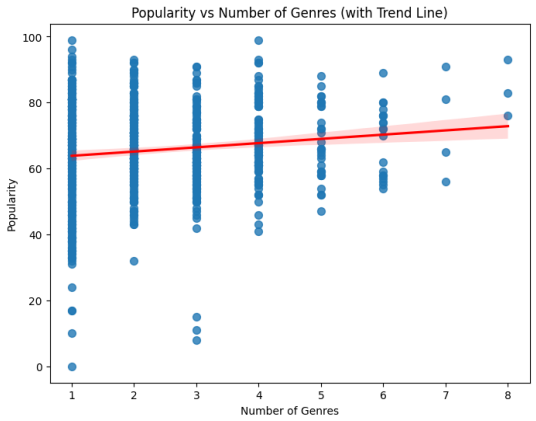

Figure 5: Primary Data Scatterplot

The scatter plot shows **Popularity** (0-100) against the **Number of Genres** (1-8). While popularity scores are widely spread (from near **0** to **100**) for any number of genres, the trend line indicates a very slight positive correlation. Popularity increases minimally, from about **64** for **1** genre to approximately **74** for **8** genres.

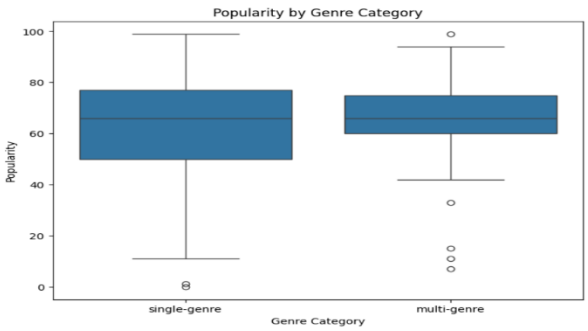

Figure 6: Primary Data Boxplot

This box plot compares popularity between "single-genre" and "multi-genre" categories. The median popularity is nearly identical: about **66** for single-genre and **65** for multi-genre. The middle **50%** of single-genre data spans roughly **50** to **78**in popularity, while the middle **50%** of multi-genre data spans approximately **60** to **75**.

Insights:

1. Multi-genre artists tend to have higher popularity, suggesting that genre diversity broadens audience reach.
2. Genre flexibility may improve cross-regional success.
3. Future hypothesis: Artists collaborating across genres may achieve higher global popularity and greater chart stability

Secondary Data :

Contextual Information

The secondary dataset was obtained from Kaggle, where it was published as a cleaned and aggregated version of Spotify data. This dataset provides artist popularity and genre information collected and shared by other researchers for public use.

Metadata Review

Source: The dataset was obtained from Kaggle, where it was published by contributors who collected and prepared Spotify artist data.
Date Collected: The dataset was published on Kaggle in late 2025 (last updated ~May 2025).
Collection Method: Data was originally extracted from Spotify's API by the Kaggle contributor, then cleaned, aggregated, and shared as a ready-to-use CSV file.
Content: The original dataset includes artist's name, artist's genre, artist's followers, artist's popularity, artist spotify URL, and track details such as the name, album name, release date, duration, explicit flag, track popularity, as well as audio features such as Danceability, Energy, Key, Loudness, Mode, Speechiness, Acousticness, Instrumentalness, Liveness, Valence, Tempo. However track and audio features are not relevant and therefore we dropped those columns and additional derived features like genre count and category were computed during preprocessing.

Bias Awareness

- Population Bias: The dataset represents Spotify users only; listening behavior may not reflect global music consumption.
- Temporal Bias: The snapshot may not match current Spotify values since artist popularity changes over time.
- Collection Bias: Because the dataset was uploaded by a Kaggle user, preprocessing choices (e.g., how missing values or duplicates were handled) may influence results.

EDA Results

Data Cleaning & Preparation

- Standardized column names (track_name, artist_name, streams, region, date).
- Converted types (streams → numeric, date → datetime).
- Removed duplicate rows and handled missing values.
- Derived features where relevant (genre_count, genre_category).

Descriptive Statistics by Genre Category

Max	75%	50% (Median)	25%	Min	Std	Mean Popularity	Count	Genre Category
91	82.5	67.0	72.0	49	7.75	76.78	67	Multi-genre
92	75.5	77	0	0	37.04	34.73	51	Single-genre

Interpretation:

The table shows that multi-genre artists are generally more popular than single-genre artists, with a higher mean (76.78) and median (72.0), as well as a narrower spread (std = 7.75) — indicating consistent popularity across this group.
In contrast, single-genre artists have a much lower average popularity (34.73) and greater variability (std = 37.04), suggesting that while some are highly popular, many have low popularity scores.

Visual Findings (Secondary Data)

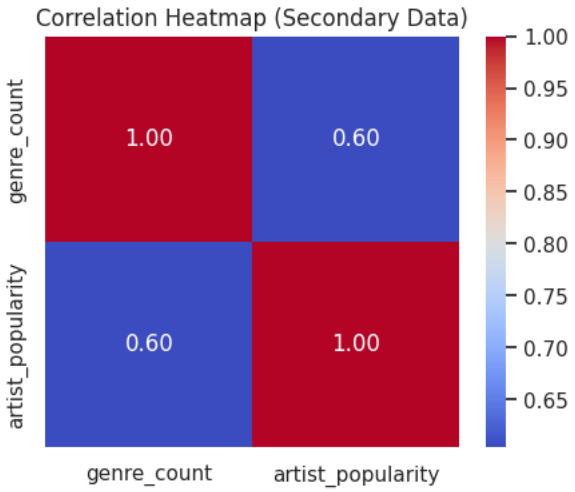

Figure 7: Secondary Data Heatmap

Correlation Heatmap

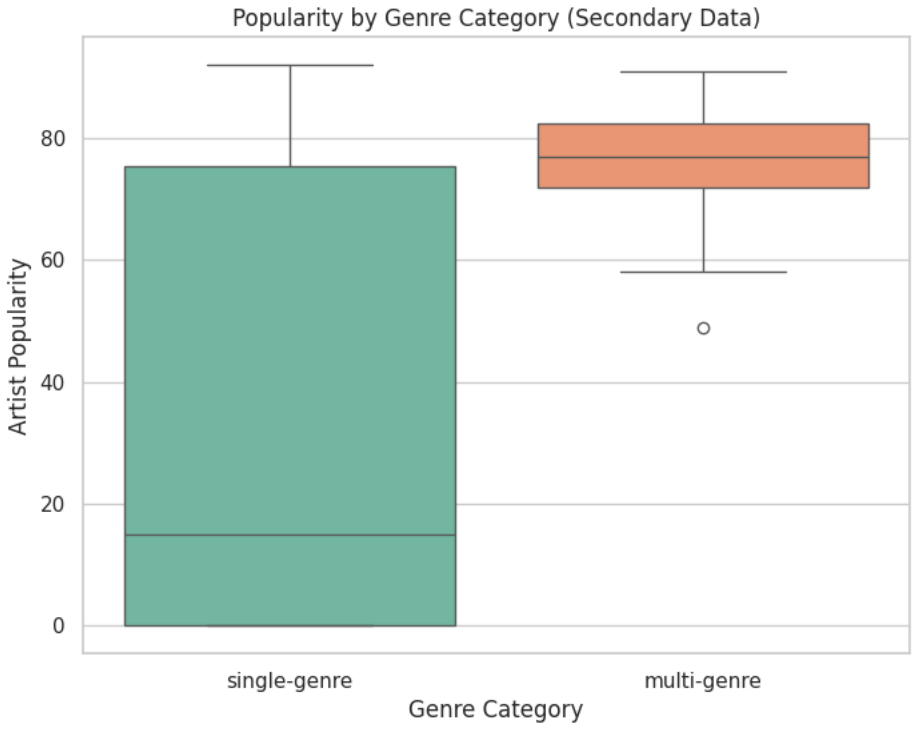

Figure 8: Secondary Data Boxplot

Boxplot – Popularity by Genre Category

The boxplot indicates that multi-genre artists consistently achieve higher popularity compared to single-genre artists. Single-genre artists show wider variability and more outliers, including very low popularity scores.

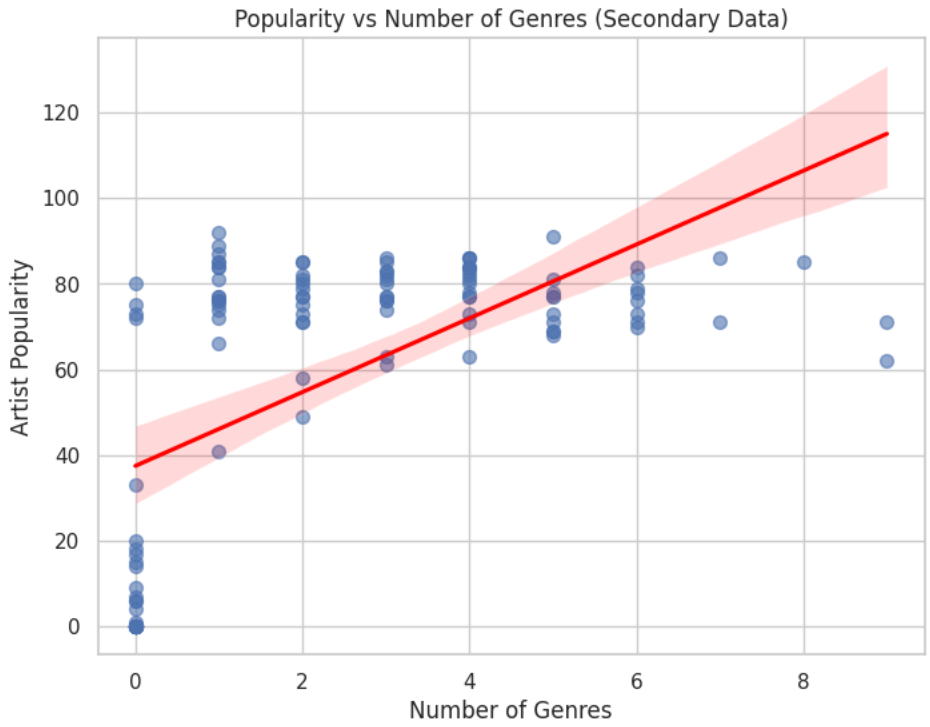

Figure 9: Secondary Data Scatterplot

Scatter Plot – Popularity vs Number of Genres

The scatter plot displays a clear upward trend: artists with more genres tend to achieve higher popularity. This aligns with the correlation finding, reinforcing the idea that genre diversity broadens audience reach.

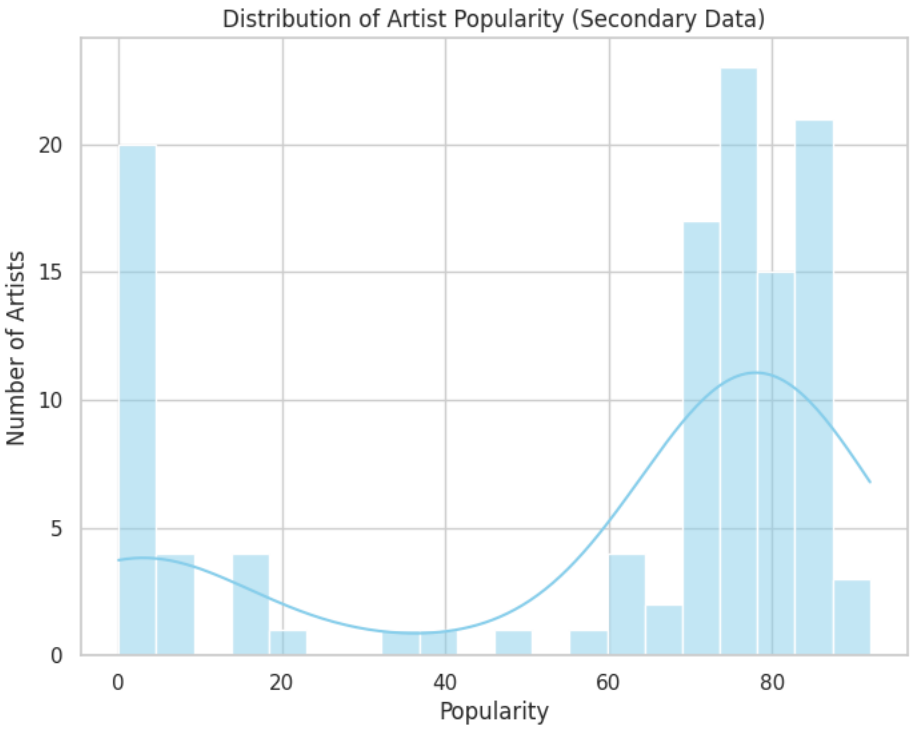

Figure 10: Secondary Data Histogram

Histogram – Distribution of Artist Popularity

The histogram shows that popularity is **right-skewed**, with most artists clustered between 60–80, but a notable group of less popular artists appears around very low values. This highlights inequality in popularity distribution.

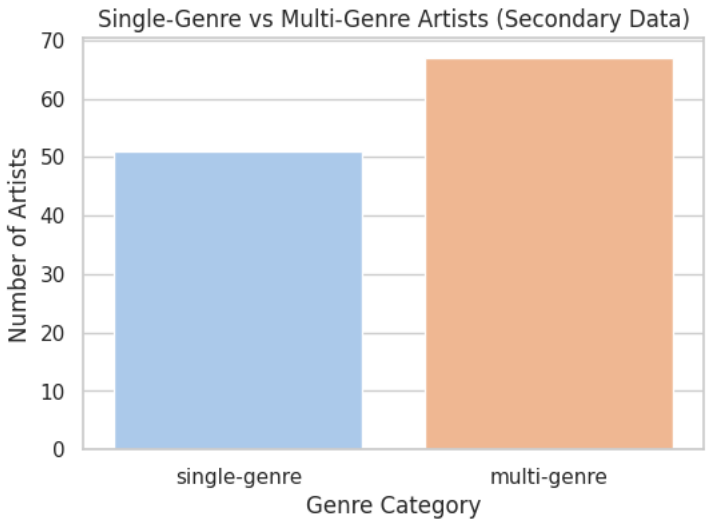

Figure 11: Secondary Data Barchart

Bar Chart – Number of Single vs Multi-Genre Artists

This bar chart demonstrates that multi-genre artists make up the majority of the dataset. This imbalance should be considered when interpreting popularity differences.

Insights:

1. Multi-genre artists dominate the dataset both in numbers and popularity.
2. The correlation between genre count and popularity is stronger in the secondary dataset than in the primary one.
3. Popularity is unevenly distributed, with a small set of highly popular artists and many with low exposure.

Comparison of the results of both datasets

Metric	*Primary Dataset*	*Secondary Dataset*	*Interpretation*
Multi-Genre Count	475	67	The secondary dataset has fewer total artists, but the proportion of multi-genre artists remains dominant in both datasets.
Single-genre count	233	51	The smaller secondary dataset maintains a similar ratio between single- and multi-genre artists.
Mean Popularity (Multi-genre)	62.04	76.78	Multi-genre artists are considerably more popular in the secondary dataset, suggesting that genre diversity has a stronger impact in newer or broader samples.
Mean Popularity (Single genre)	58.01	34.73	Single-genre artists have much lower average popularity in the secondary dataset, reinforcing the advantage of cross-genre appeal.
Median Popularity (Multi Genre)	61	77	The median shows a higher baseline for multi-genre artists in the secondary dataset, indicating more consistent performance.
Median Popularity (Single genre)	61	15	The drop in median popularity for single-genre artists in the secondary data highlights declining visibility or niche concentration.
Standard Deviation (Multi Genre)	12.75	7.75	Popularity among multi-genre artists is more stable in the secondary dataset.
Standard Deviation (Single Genre)	18.75	37.04	Popularity among single-genre artists is highly volatile, showing extreme variation in success.
Range (Multi Genre)	5-95	49-91	Both datasets show similar popularity ranges, but the secondary dataset lacks very low performers.
Range (Single Genre)	0-95	0-92	Wide range persists, but more extreme low scores dominate in the secondary dataset.

In both datasets, multi-genre artists consistently outperform single-genre artists. The secondary dataset amplifies this trend — the difference in mean popularity between multi- and single-genre artists is nearly double that observed in the primary dataset.

Contextualizing Findings

The secondary dataset provides a stronger, more current validation of the trend found in the primary data that artists engaging with multiple genres achieve higher and more consistent popularity.

Possible causes of differences could be the sample size. Despite the secondary dataset including fewer artists, it might cover a broader range of markets and listening demographics as well as data source variance in which primary data was collected from the Spotify API, whereas the secondary dataset came from a compiled source.

Both datasets confirm that genre diversity positively impacts artist popularity, but the secondary dataset strengthens this conclusion, showing a sharper divide and more stable multi-genre performance.

Summary of New Insights and Hypotheses

New Insights

- Multi-genre artists tend to have higher popularity, suggesting that genre diversity broadens audience reach.
- Genre flexibility may improve cross-regional success.
- In the secondary dataset, multi-genre artists dominate both in numbers and popularity.
- The correlation between genre count and popularity is stronger in the secondary dataset than in the primary one.
- Popularity is unevenly distributed, with a small set of highly popular artists and many with low exposure.

Hypotheses: Artists collaborating across genres may achieve higher global popularity and greater chart stability.

Raw Data File(s)

- thttps://github.com/aljoharas/datascience/tree/main/charts

A Jupyter Notebook

- https://github.com/aljoharas/datascience/blob/main/spotifyproject.ipynb