Exploring the Musical Landscape With Discogs

Group 5:
Abhiraam Thadur, Nursultan Azhimuratov,
Faris Hazim, Imran Iskander, Amy Cai

Analysis Roadmap

- **Introduction**: Dataset Background and Motivation
- **Data Setup**: Splitting, Parallel Computation with CHTC, Merging, and Analysis
- **Exploratory Data Analysis Results**: Trends, Correlations, Hypothesis Testing, and Machine Learning
- **Conclusion**: Summary of Findings and Future Direction

INTRODUCTION

Background & Motivation

Discogs:

- Wikipedia for music in all formats (CD, Vinyl, Digital, etc.)
- Updated every month

Dataset:

- April 2025 Snapshot – releases.csv
- Over 18 million release information → ~32 GB in size

Motivation:

- Find interesting trends and patterns (e.g. releases over time, genre popularity, etc.)
- Conduct hypothesis testing to understand genre differences
- Run a survival analysis to identify when artists remaster their original release

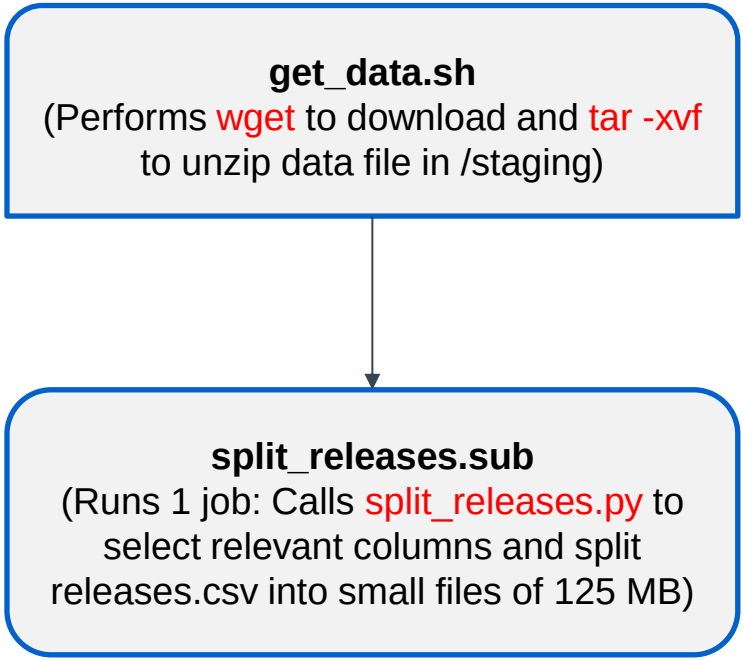
Important Variables	Description
Artist ID/Name	Main artist(s) credited
Label ID/Name	Label the artist(s) is/are signed to
Genre	High-level music category
Style	Sub-genre
Format	Release medium (e.g. Vinyl, CD, Digital)
Track List/Duration	List/Duration of tracks on the release
Release Date	Release date
Country	Country of release

DATA SETUP (DONE IN HTCONDOR)

1. Initial Download and Splitting

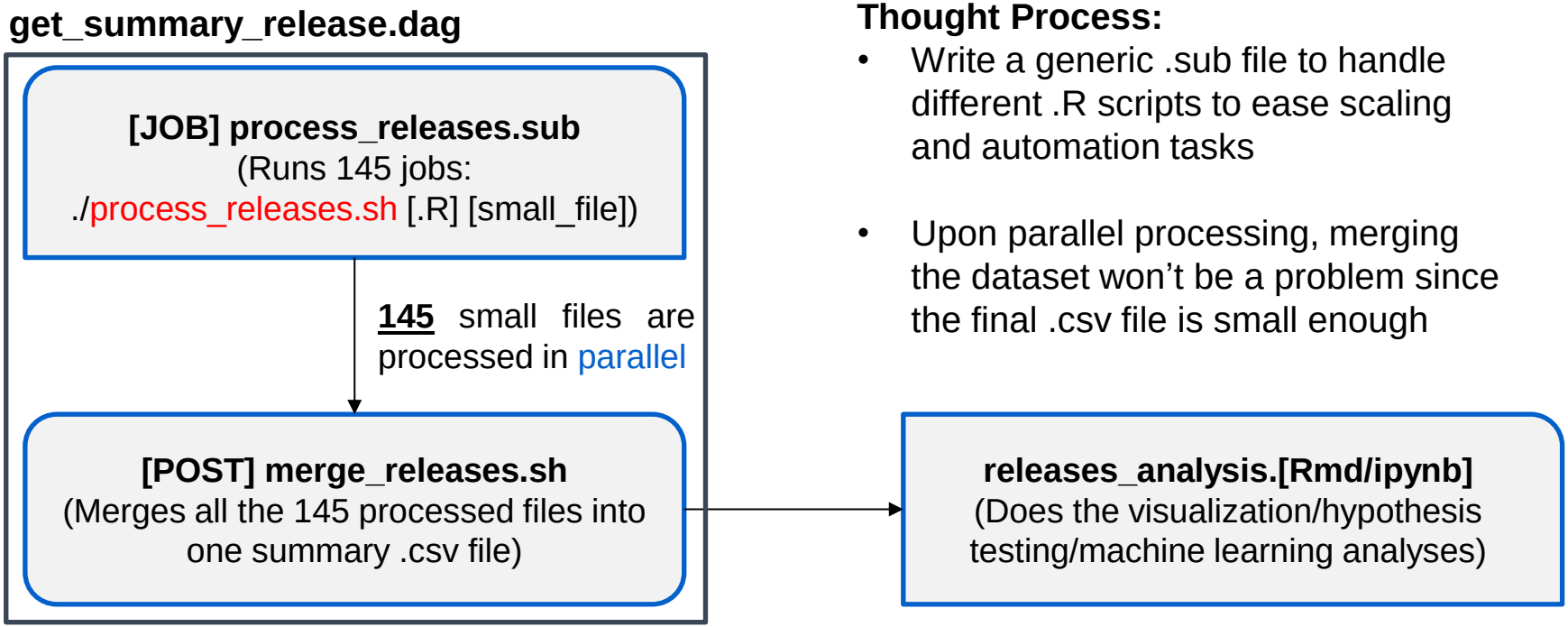
Thought Process:

- Utilize the 200 GB quota in /staging to store releases.csv and its data splits
- All statistical computations (running .sh/R scripts and submitting condor jobs) to be done within the home directory
- Use Python to split since certain columns store data in .json



Result: **145** small releases files

2. Parallel Computation and Analysis

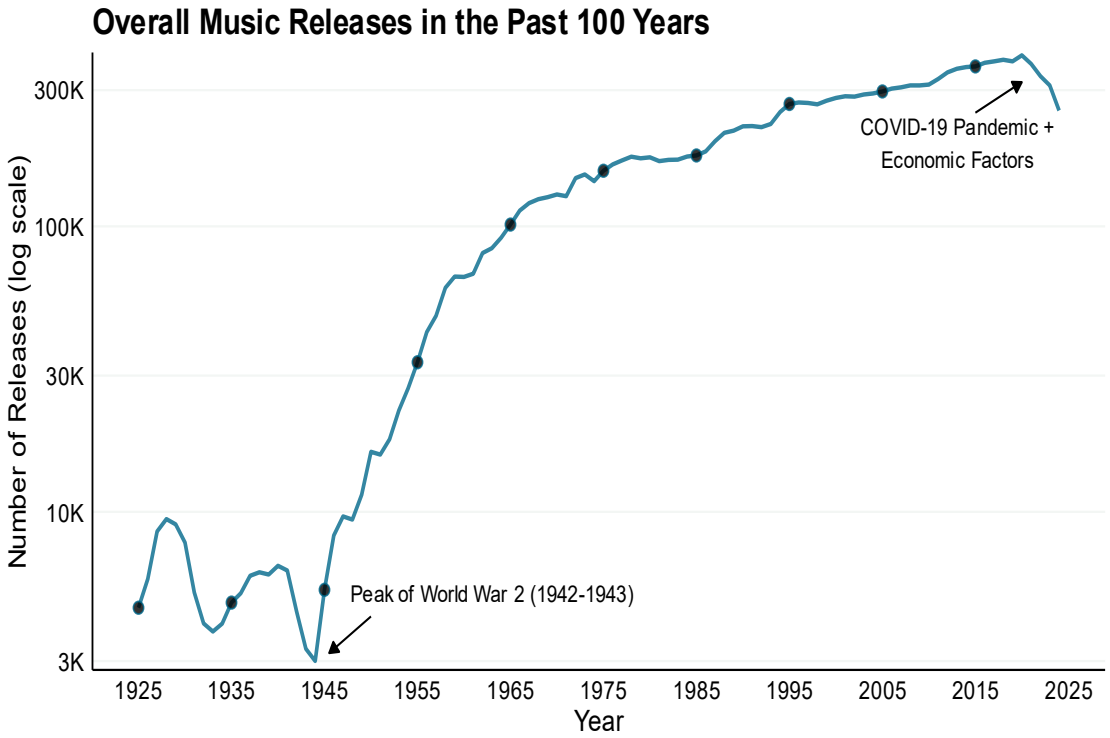


Thought Process:

- Write a generic .sub file to handle different .R scripts to ease scaling and automation tasks
- Upon parallel processing, merging the dataset won't be a problem since the final .csv file is small enough

EDA RESULTS

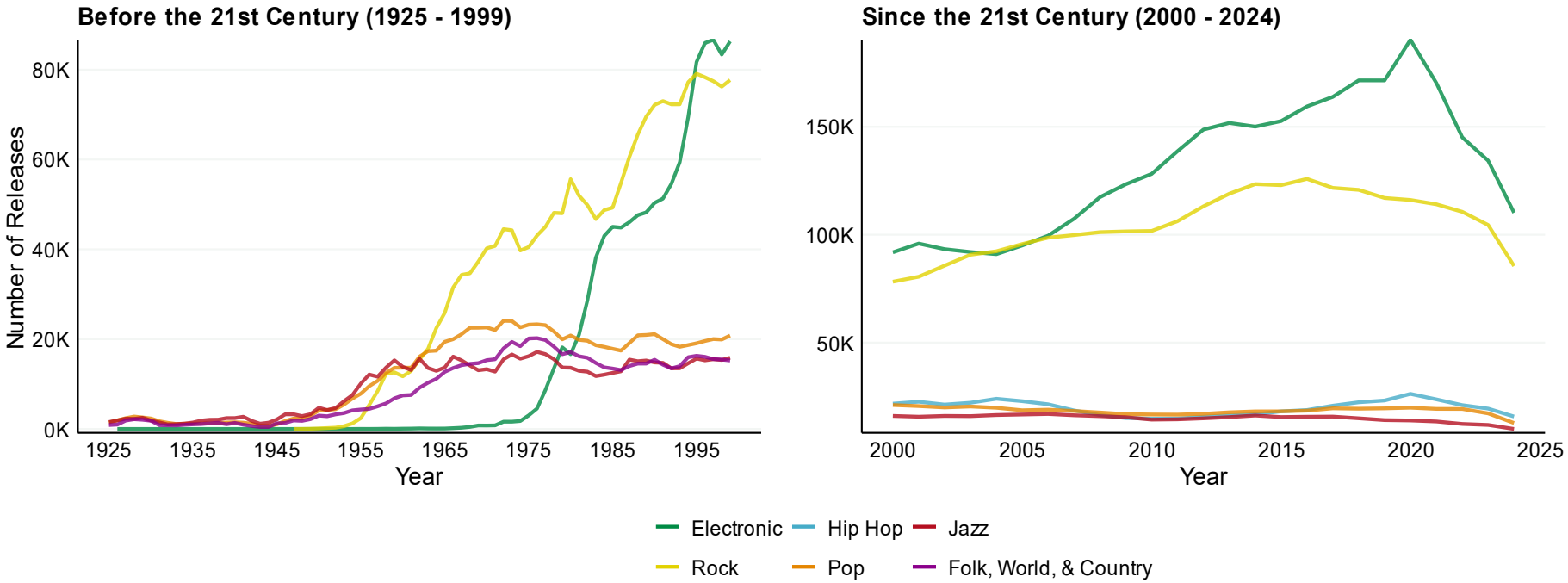
Temporal Trends



- Music releases show two significant **dips**: circa. 1942-1943 and the start of the COVID-19 pandemic
- The first dip: Musicians went on strike against record labels for unfair wages
- The second dip: economic shutdown, shift in consumer behavior, and album releases no longer bringing in huge profit

Temporal Trends

Evolution of Music Genre Popularity

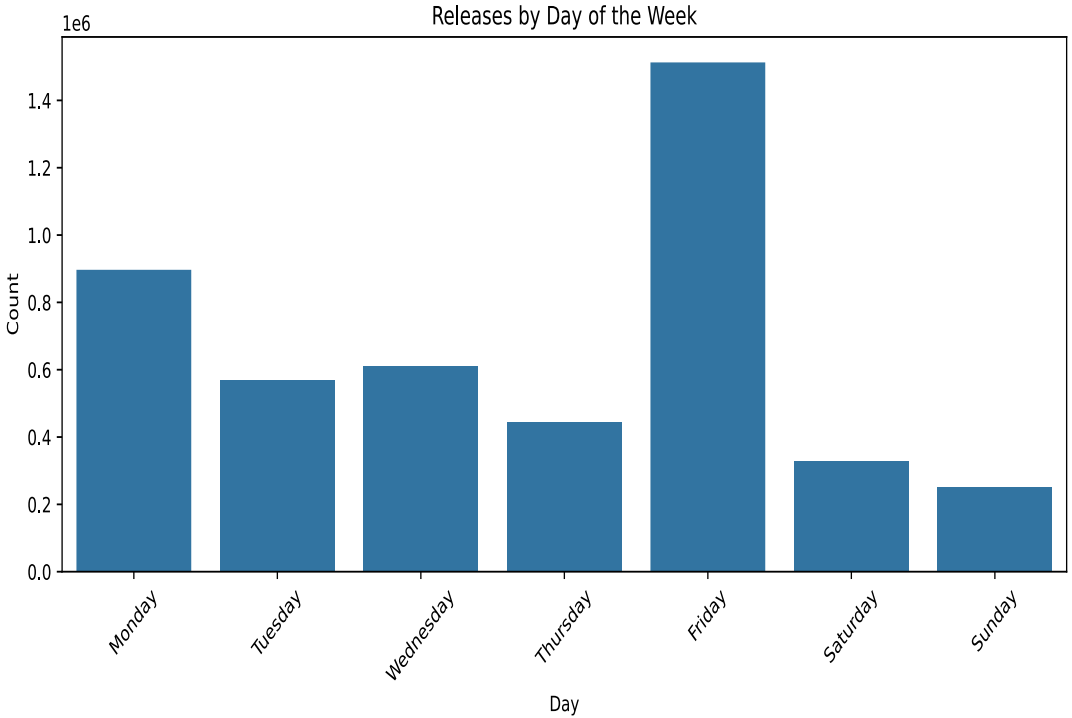


Temporal Trends

Before: Releases were staggered by market—UK on Mondays, US on Tuesdays, Japan on Wednesdays, and other territories on varying weekdays

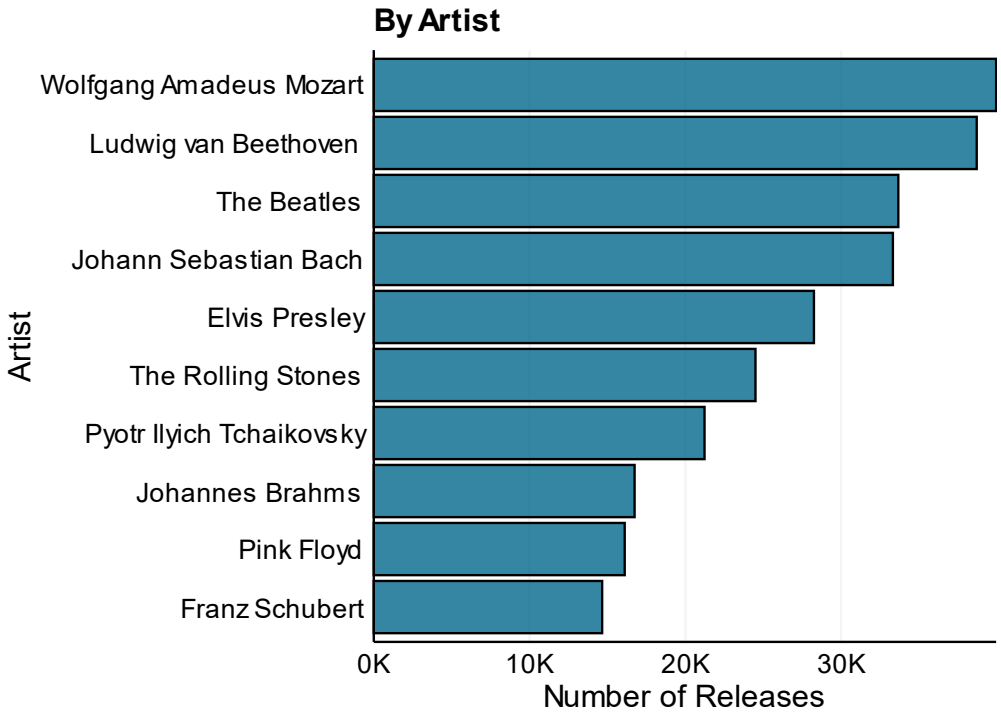
Friday: Official Global Release Day (since 2015), ensuring simultaneous worldwide launches

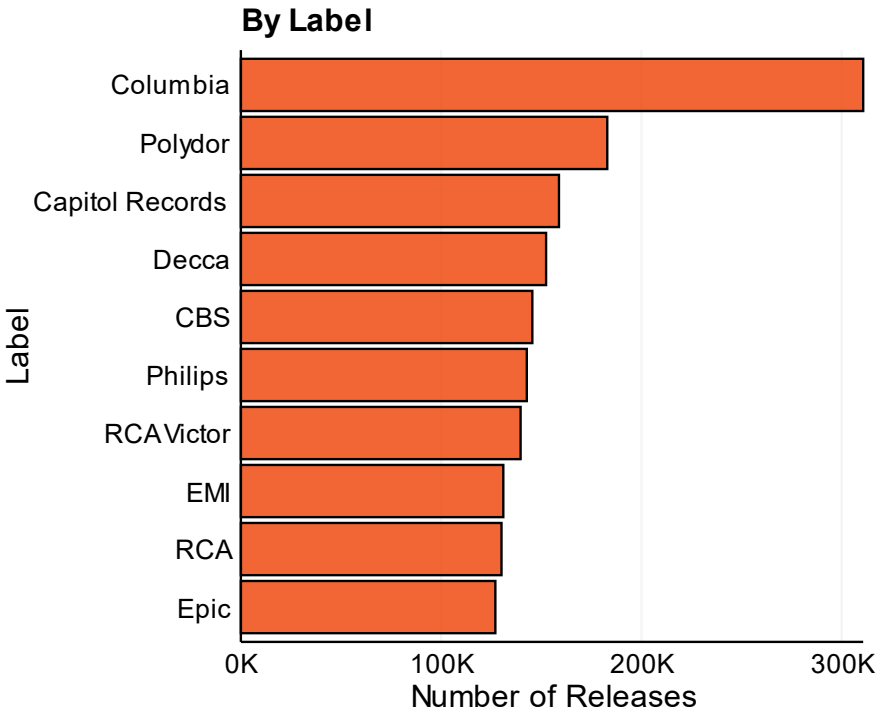
Purpose: Focus listener attention on weekends, unify chart tracking windows, streamline global marketing, and curb early leaks and piracy



Top Releases

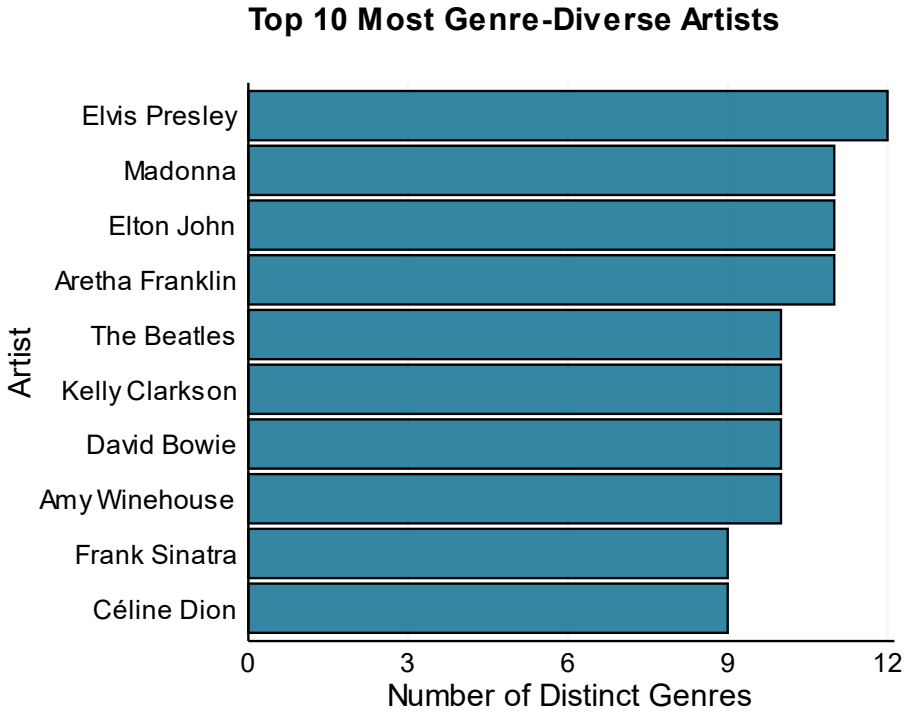
Artists and Labels with the Most Releases

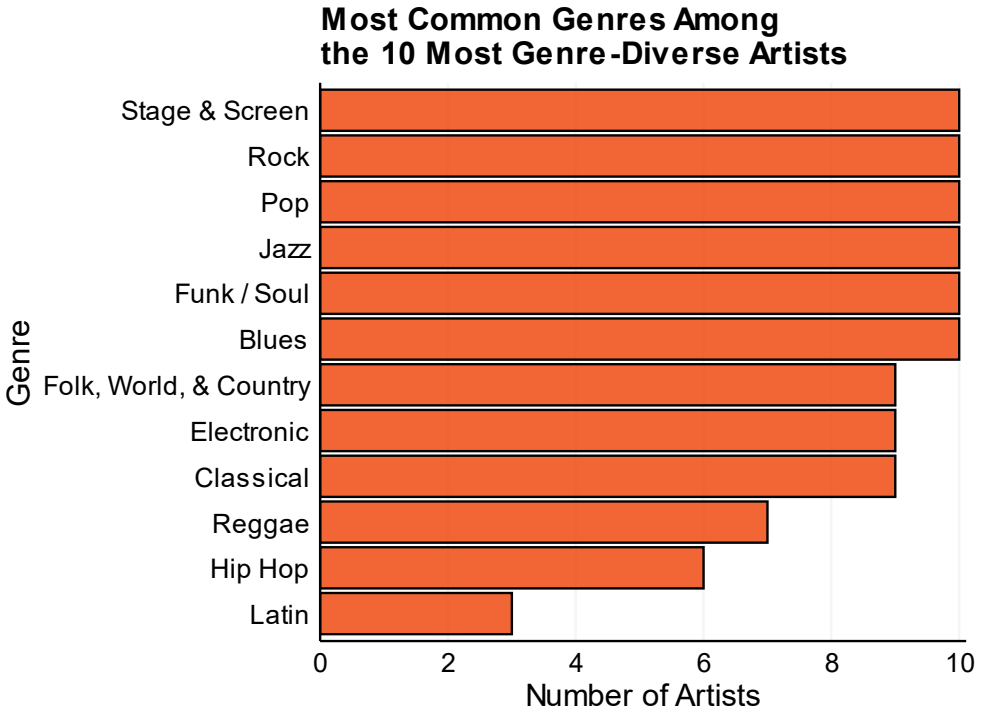




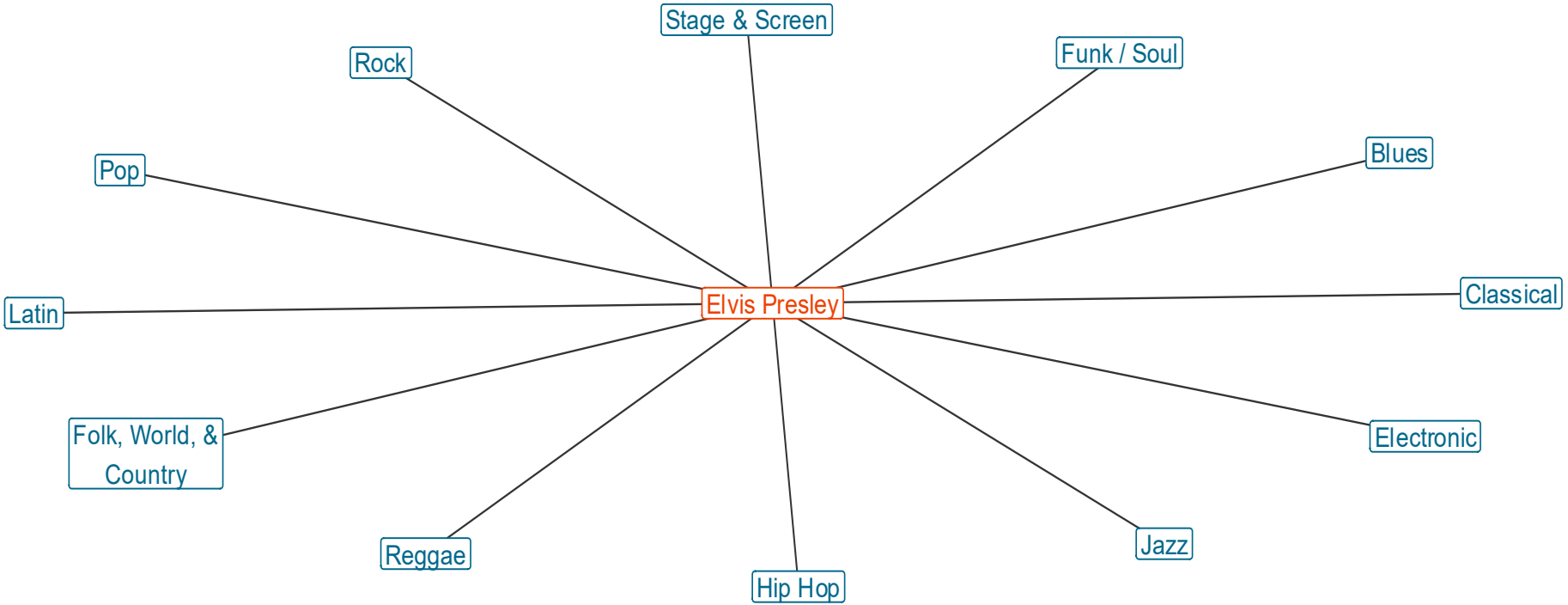
Genre-Diversity

Genre Diversity (All Time)

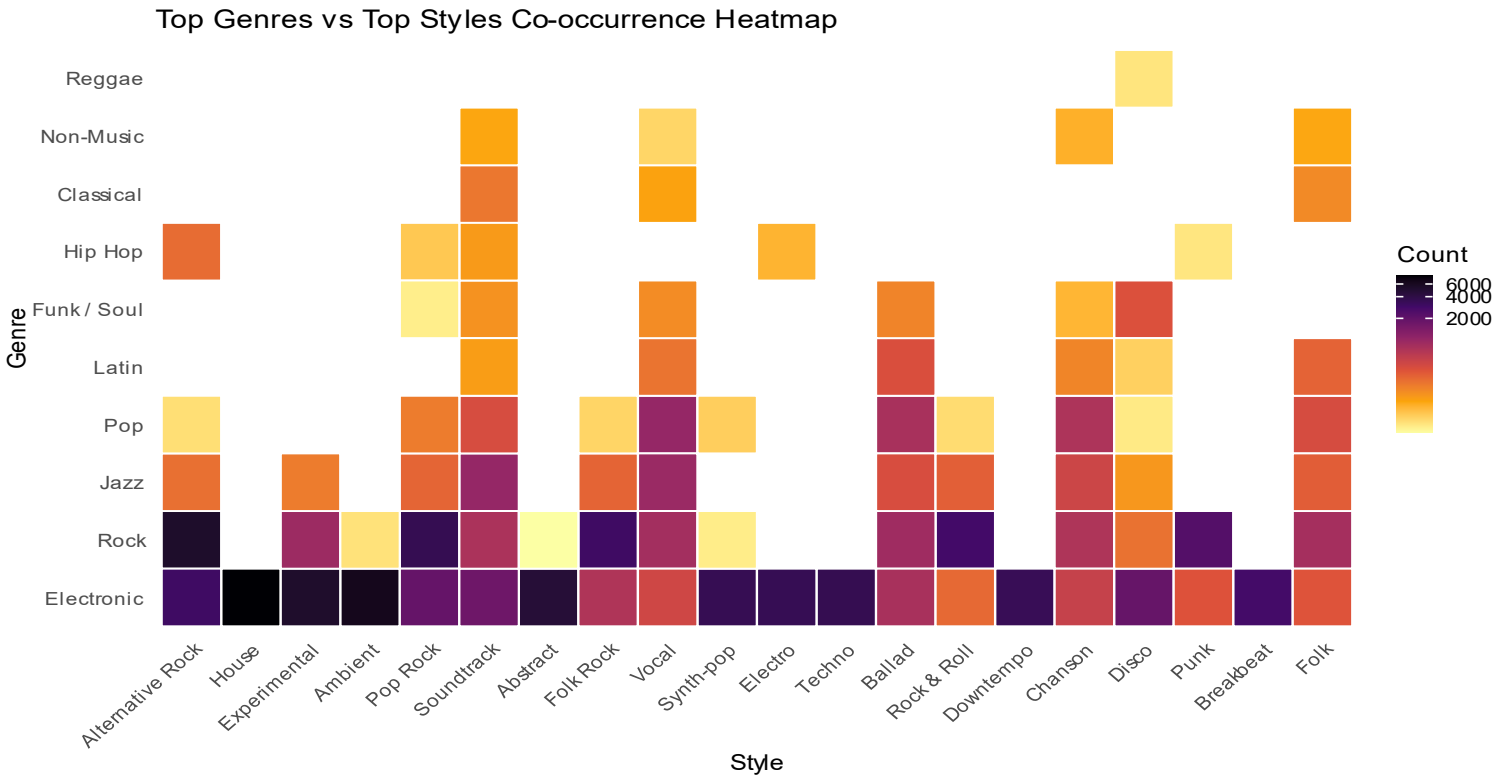




Elvis Presley and the Genres He was Associated With



Genre-Style Association



Hypothesis Testing (One-Way ANOVA)

Null Hypothesis (H_0): $\mu_{Blues} = \mu_{Jazz} = \ldots = \mu_{Rock}$

Alternative Hypothesis (H_1): Not all μ_i are equal

```
##                   Df    Sum Sq   Mean Sq F value Pr(>F)
## genre             14 1.046e+11 7.469e+09   78562 <2e-16 ***
## Residuals   10143506 9.644e+11 9.507e+04
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

$p < 2 \times 10^{16}$, much less than 0.05

Since $p < 0.05$, H_0 was rejected, indicating that **differences in mean duration do exist** between genres.

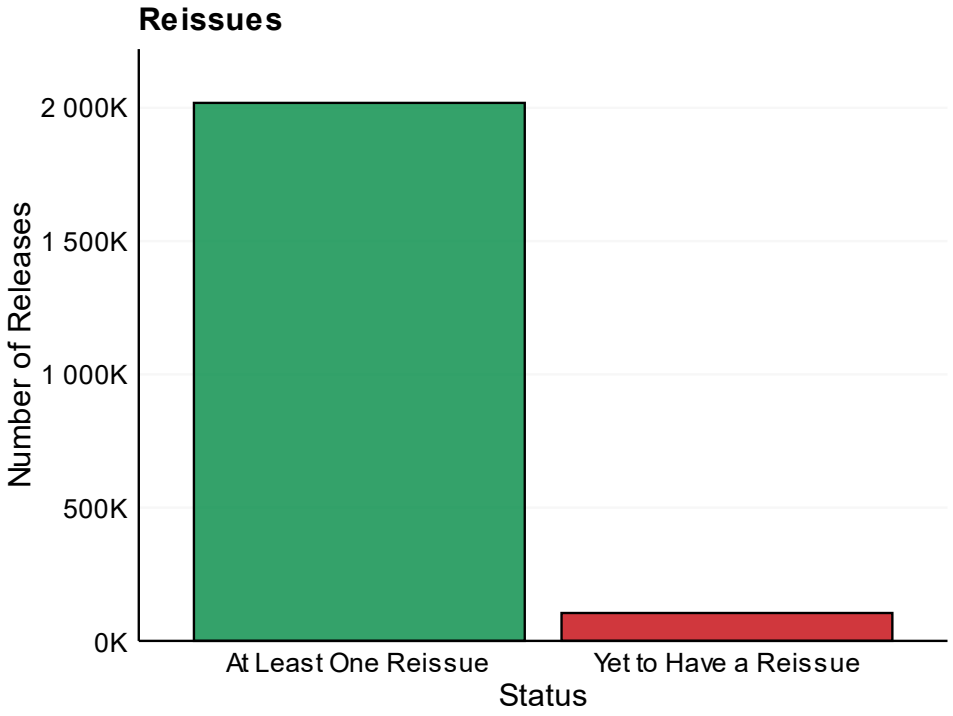
Significantly different pairs ($p < 0.05$)	96
Non-significantly different pairs ($p \geq 0.05$)	9
Largest mean difference	Non-Music vs. Blues ($p < 0.001$)
Smallest non-significant difference	Folk vs. Blues ($p = 0.9958$)

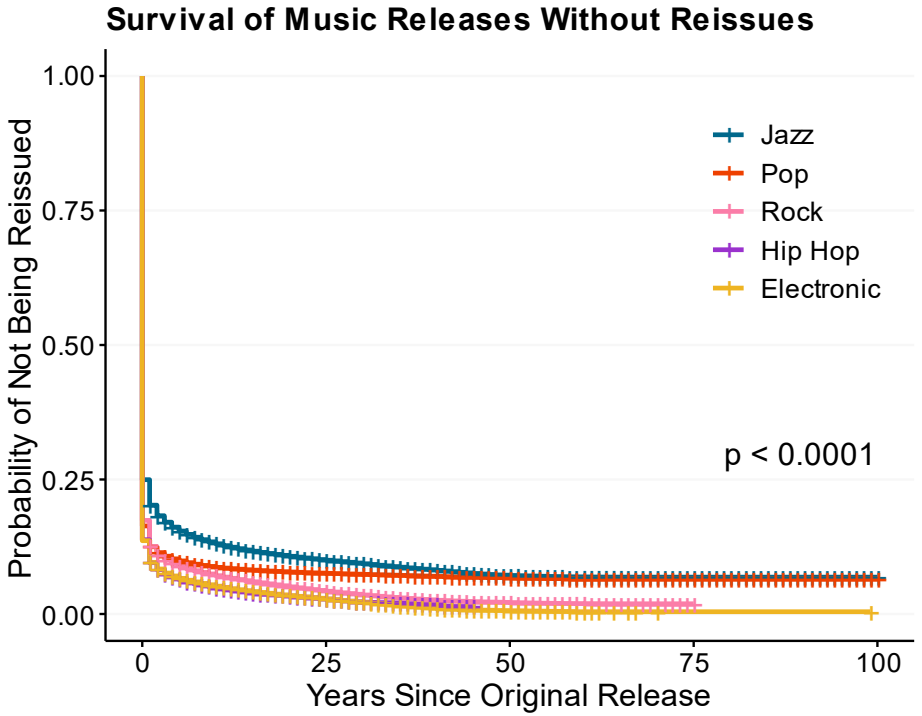
Genre	Similar genres (no significant difference)
Blues	Brass & Military; Folk; Funk / Soul
Brass & Military	Blues; Folk; Latin; Pop; Stage & Screen
Folk	Blues; Brass & Military
Funk / Soul	Blues
Latin	Brass & Military
Pop	Brass & Military
Stage & Screen	Brass & Military

Genres different from all others	Children’s; Classical; Electronic; Hip Hop; Jazz; Non-Music; Reggae; Rock

Survival Analysis

Do Artists Remaster Their Original Release ?





CONCLUSION

Discussion and Future Direction

Key Points:

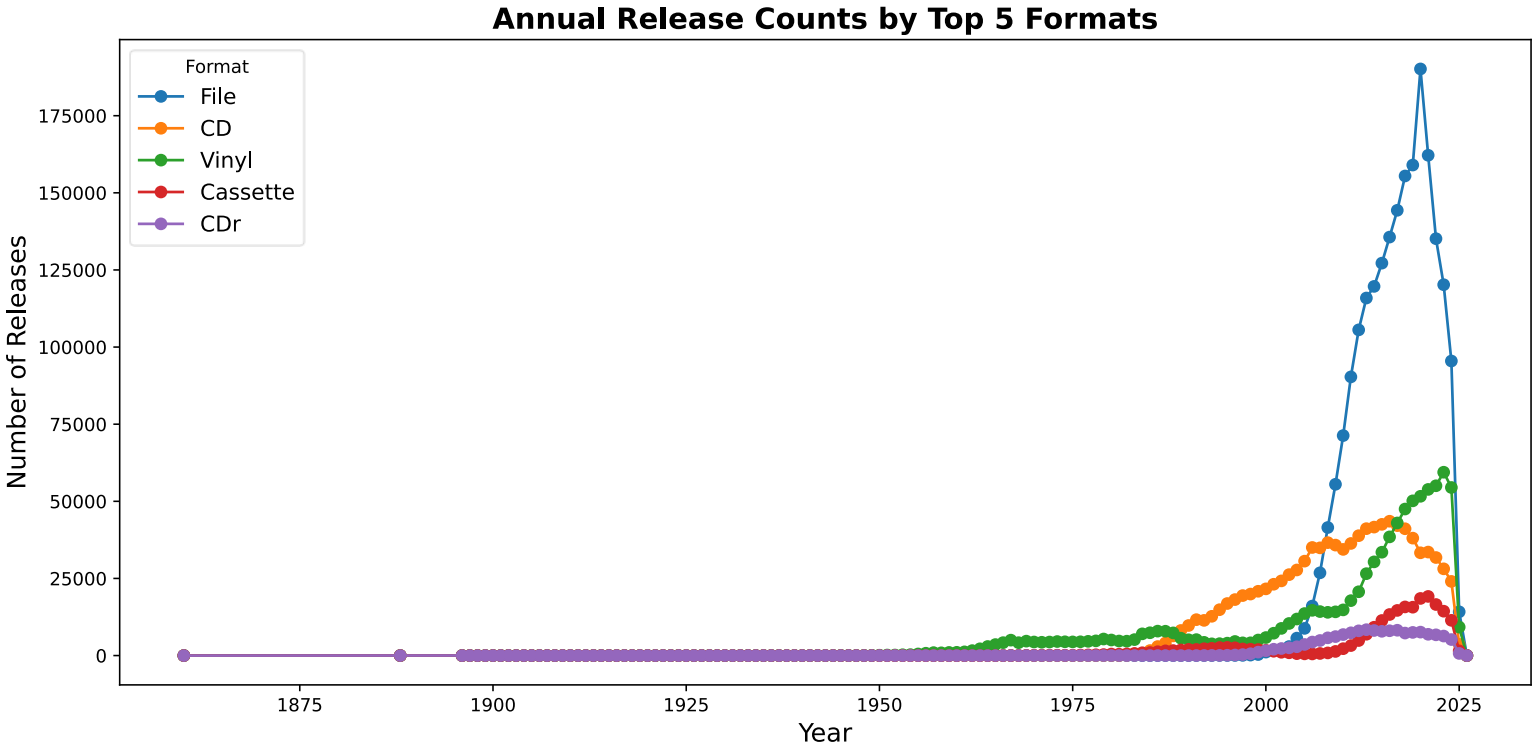
- Historical Disruptions on music release during world war and COVID-19
- Global release on Friday reduce piracy and increase listener
- Elvis Presley is the most versatile artist
- Electronic is the genre with the most genre-style combination
- There is a difference in mean duration for music in different genre
- Electronic songs/albums usually re-release very quickly

Future:

- Integrate streaming numbers from Spotify data to quantify song popularity over time
- Text mining on release title to identify any language shift across time

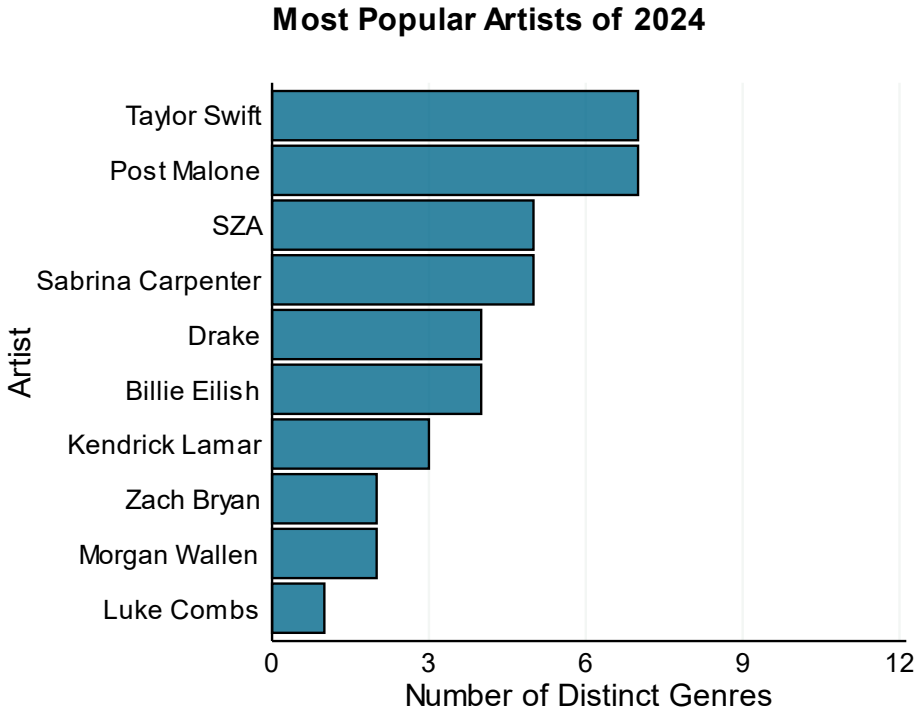
APPENDIX (EXTRA ANALYSIS)

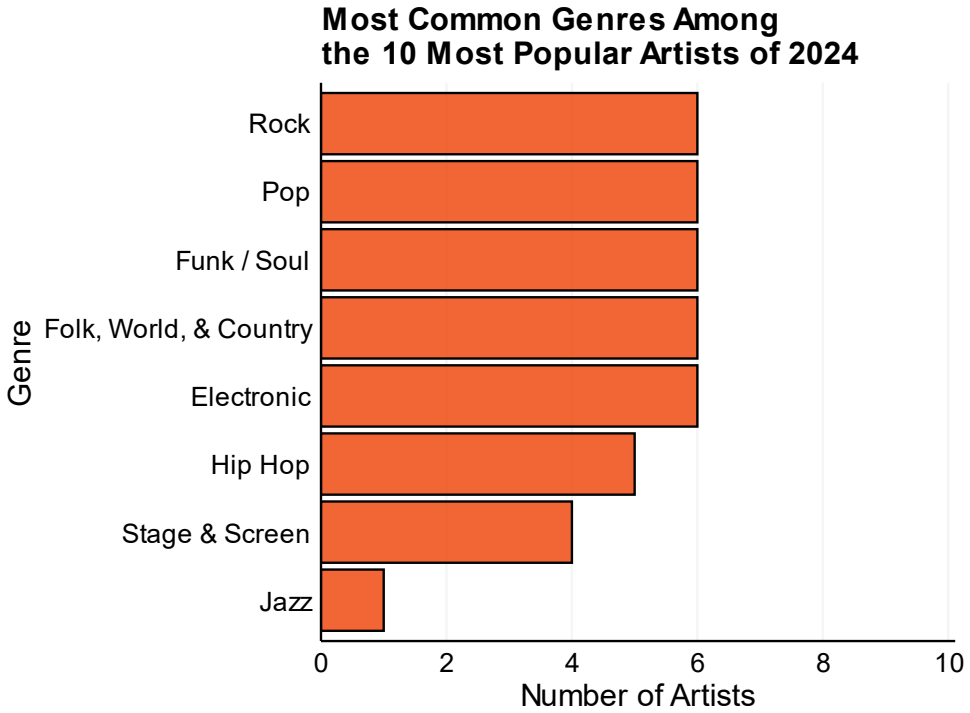
Temporal Trends



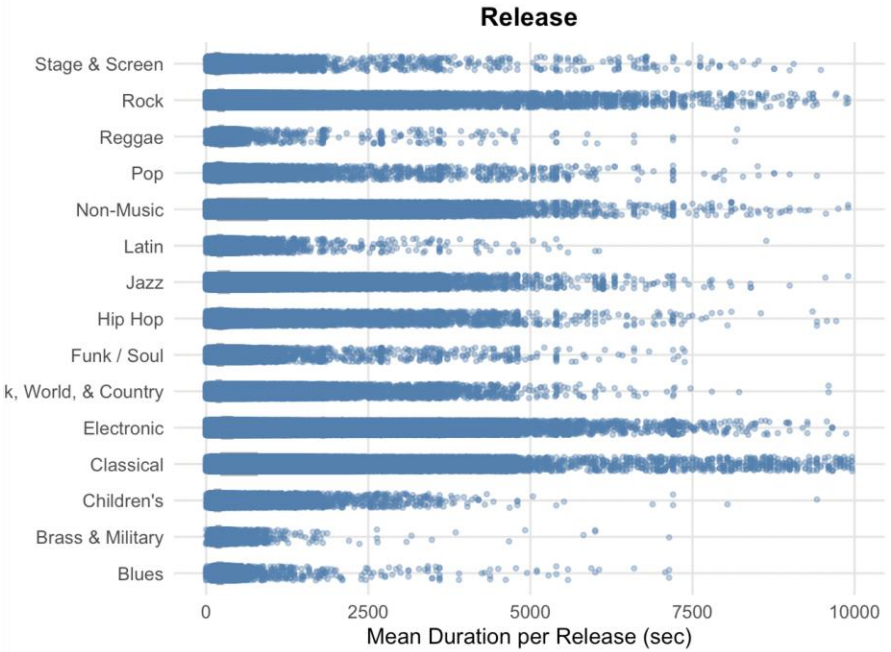
Genre-Diversity

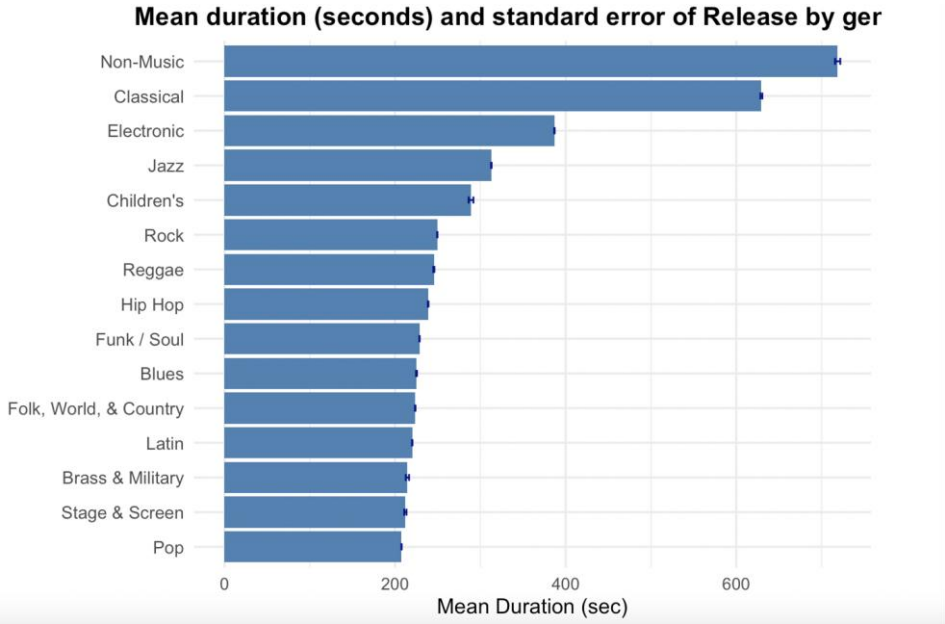
Genre Diversity (Current)



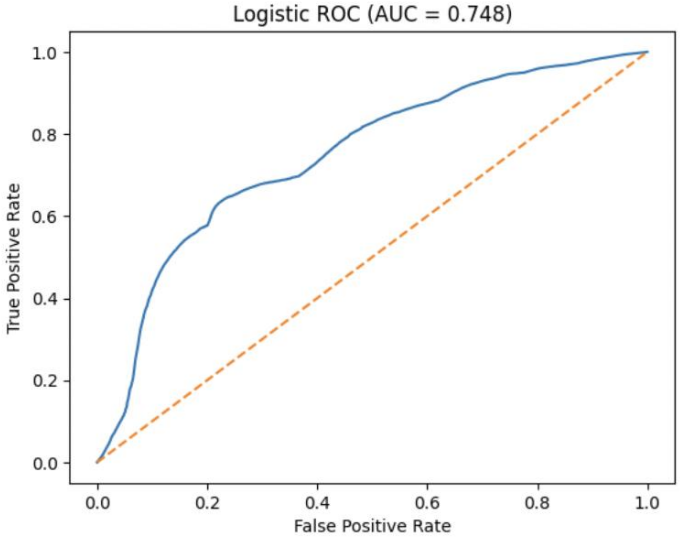


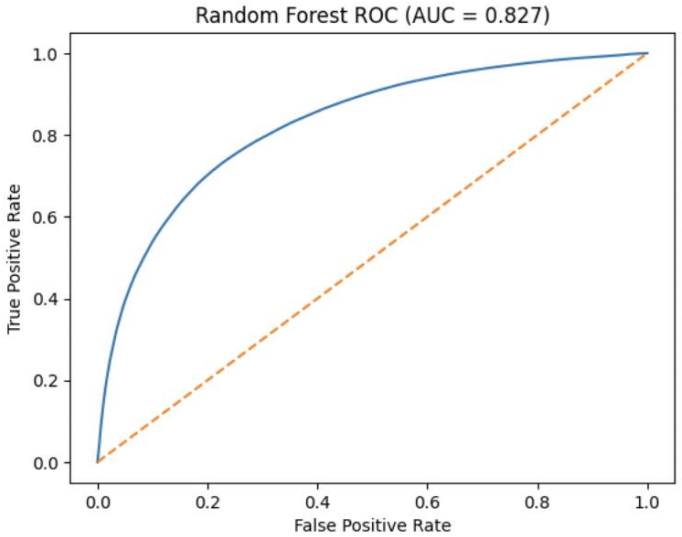
Hypothesis Testing (One-Way ANOVA)





Predicting a Song's Mean Duration





X：
N_track
Num_extra_artists
Release_year
genre_... (15 columns of one-hot coding)

y:
A binary label with a value of 0 or 1
1: the average length of this posting exceeds a certain threshold ("long")
0: the average length does not exceed the threshold ("short")