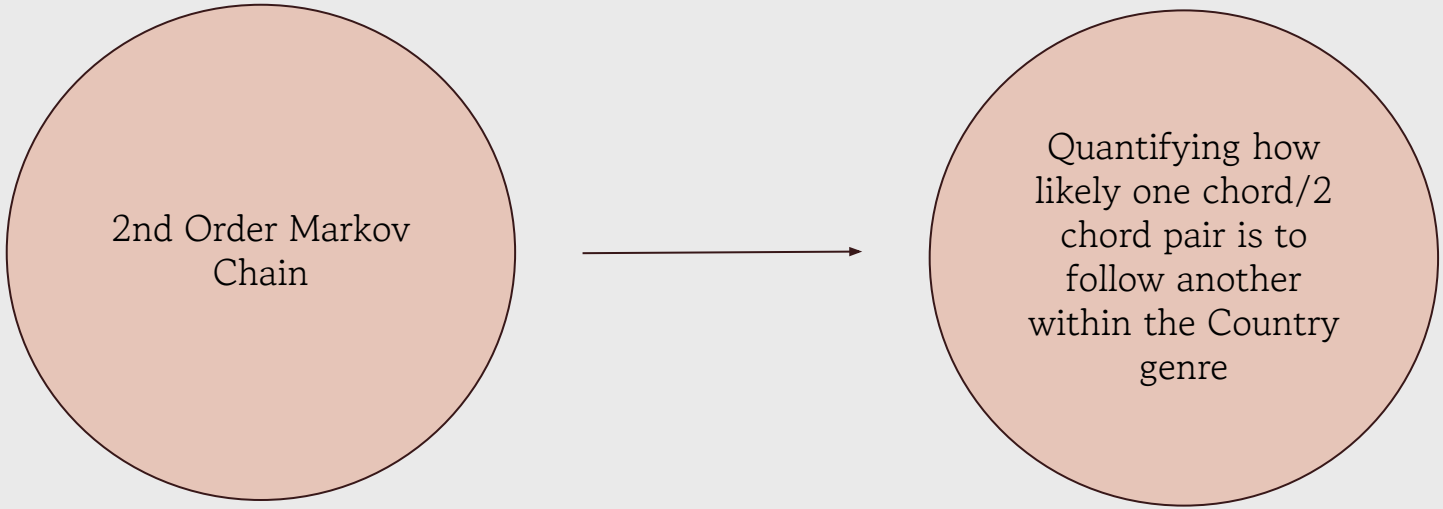
# DS 5030: Chordonomicon Group Project

Isabela Barton, Hayeon Chung, Ben Doniger

# Introduction: What We Are Modeling



# Data

- Focused on 53,306 **“country”** songs for this analysis
- Each record includes artist, genre, song identifiers, release year, and full chord progression

```
In [9]: countrydf.isna().sum()

Out[9]: id                    0
        chords                0
        release_date       8344
        genres                0
        decade             8344
        rock_genre        38223
        artist_id             0
        main_genre            0
        spotify_song_id    6059
        spotify_artist_id     0
        dtype: int64
```

**Missing Data Highlights:**
- 8,000 songs lack release dates → limits decade-by-decade trend analysis
- 38,000 songs missing sub-genre labels → harder to examine country-rock crossover trends
- 5,000 songs missing Spotify IDs → restricts linking to popularity or external metadata

Despite some missing metadata, the dataset offers a rich, large-scale view of harmonic structure in country music and serves as a strong foundation for statistical modeling.

# Methods

### 1. Data Preparation

- Filtered the Chordonomicon dataset to include 53,000 country songs only

- Cleaned inconsistent chord notation (ex. “G/B,” “Dmaj7”) and simplified slash chords to root forms for model compatibility

- Removed section tags (like <verse_1>, <chorus_1>) to isolate only chord symbols

- Verified completeness: 10 columns → key ones include *id, chords, release_date, genre, artist_id, main_genre, spotify_song_id*

### 2. Model Construction

- Implemented both first-order and second-order Markov transition models

- Treated each chord (or chord pair) as a discrete state; transitions represent conditional probabilities between chords

- Computed transition count matrix and normalized it to obtain a transition probability matrix

- 3 visualizations (Heatmap, 2-chord pair histogram, forecasting plots)

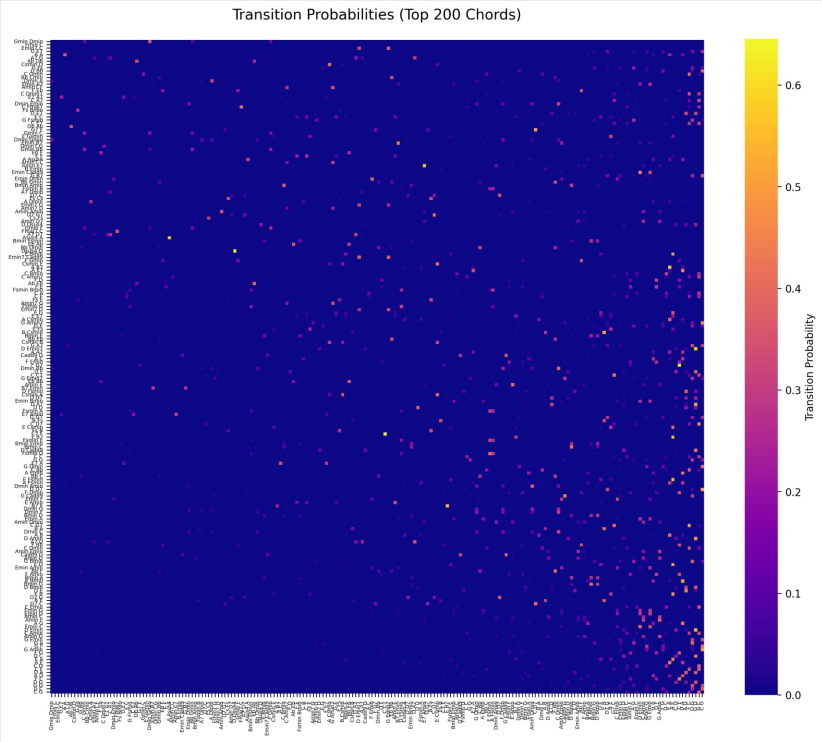
### 3. Sequence Generation

- Initialized generation with the most common chord pair (“G C”)

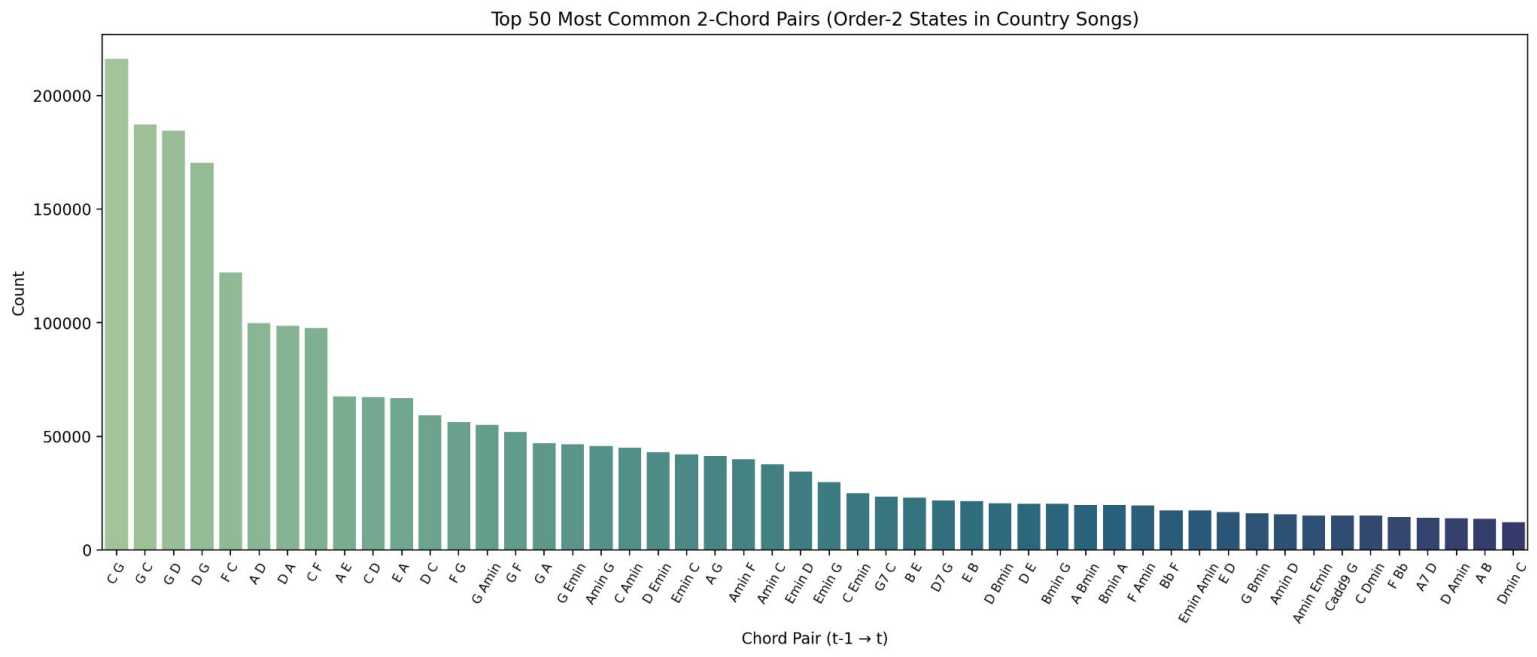
- Iteratively sampled next chords from the transition probabilities to generate realistic progressions

- Simplified chord symbols for MIDI rendering via the music21 library to enable audible playback

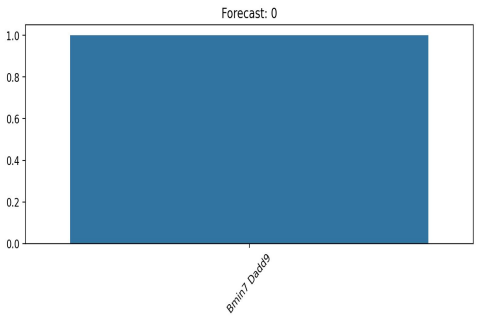
# Visualization 1: Transition Probabilities Heatmap

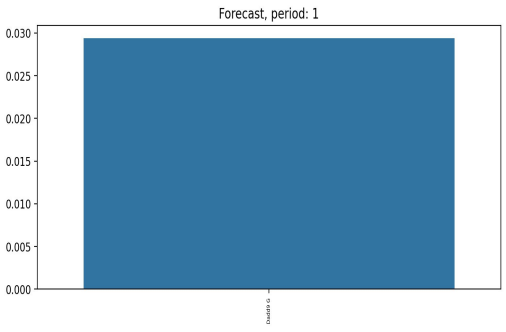


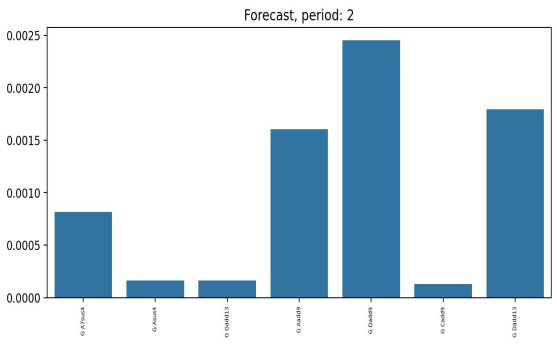
# Visualization 2: Top 50 2-Chord Pairs Histogram

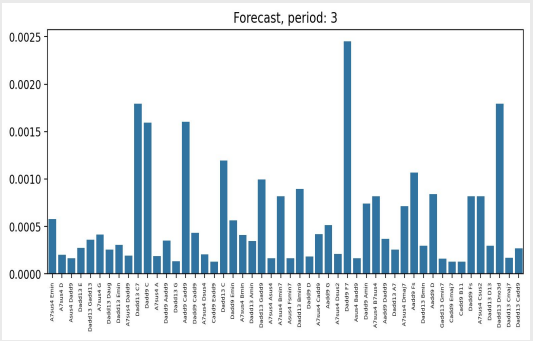


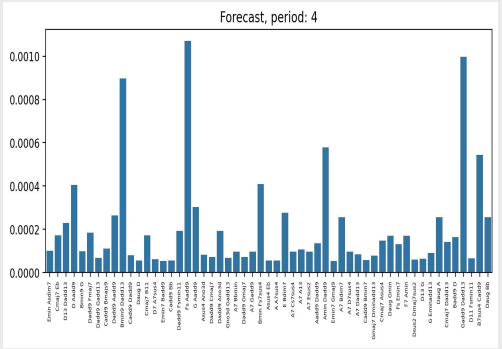
# Visualization 3: Forecasting Plots

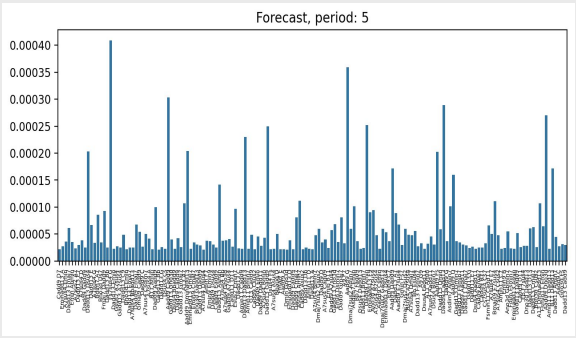


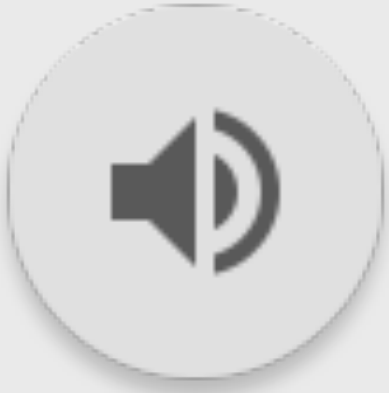



## Results



# Conclusion

**Conclusion:**

- The Markov model effectively captured short-term harmonic transitions that define country music's tone
- Results highlight a core set of dominant chord pairings that act as the harmonic framework of the genre
- Model reveals a balance between predictability and musical variation, mirroring the structure of country songs

**Limitations:**

- Fails to represent long-range song structure (verses, choruses, modulations)
- Does not account for rhythm, melody, or expressive timing
- Simplified chord notation can obscure nuanced or rare harmonic relationships

**Future Implications:**

- Extend to higher-order Markov or neural models (LSTM / Transformer) for long-term harmonic structure
- Integrate rhythmic and tonal context (e.g., key-specific transitions, chord duration)
- Analyze artist- or era-specific transition matrices to uncover stylistic diversity