Data Collection and Analysis: Comprehensive Analysis of Spotify Listening Data: April to December 2024

`Google Colab Notebook Link`

1. Introduction

This report details the findings from a detailed analysis of personal Spotify listening data, spanning April to December 2024. The primary objective was to uncover trends, patterns, and insights related to listening habits, artist preferences, song correlations, and session behavior. Data preprocessing, visualization, and statistical analysis techniques were employed to extract meaningful insights, with a strong emphasis on understanding the underlying user behavior.

2. Data Preprocessing

2.1. Data Collection

- Data was collected through an online service, Last.fm, which tracks the user's listening history on Spotify by integrating with the Spotify account.
- Last.fm gathers this information by either web scraping or through its API that provides structured listening history.
- The data used in this report was collected using the tool `lastfm-to-csv`, which enables efficient downloading of listening data in CSV format.

2.2. Source and Structure

- The dataset included four columns: `Artist`, `Album`, `Title`, and `Timestamp`.
- Data was fetched from the Last.fm/Spotify database, stored in CSV format.
- The CSV file was then uploaded to my github to fetch data into google colab `CSV File`

2.3. Timestamp Conversion

- Timestamps were converted to the **Indian Standard Time (IST)** zone for localized analysis.
- A new column, `Timestamp_IST`, was created for ease of interpretation.
- Timestamps were further decomposed into attributes such as **hour**, **day of the week**, **month**, and **date** to provide granular insights.

2.4. Data Cleaning

– Duplicates were removed, and missing values were checked. No significant issues were identified.

2.5. Derived Columns

– **Hour of Day**: Categorized into *Morning* (6–12), *Afternoon* (12–18), *Evening* (18–24), and *Late Night* (0–6).

– **Date** and **Day of Week**: Derived for time series analysis and weekly trends.

– **Session Attributes**: Calculated based on listening gaps and consecutive plays.

3. Time Series Analysis

3.1. Daily and Monthly Trends

– **Daily Listening Behavior**:

  * Activity peaked during **weekends**, particularly Saturdays, suggesting increased leisure time.
  * A sharp rise in listening frequency was noted during holidays and specific weeks in June.

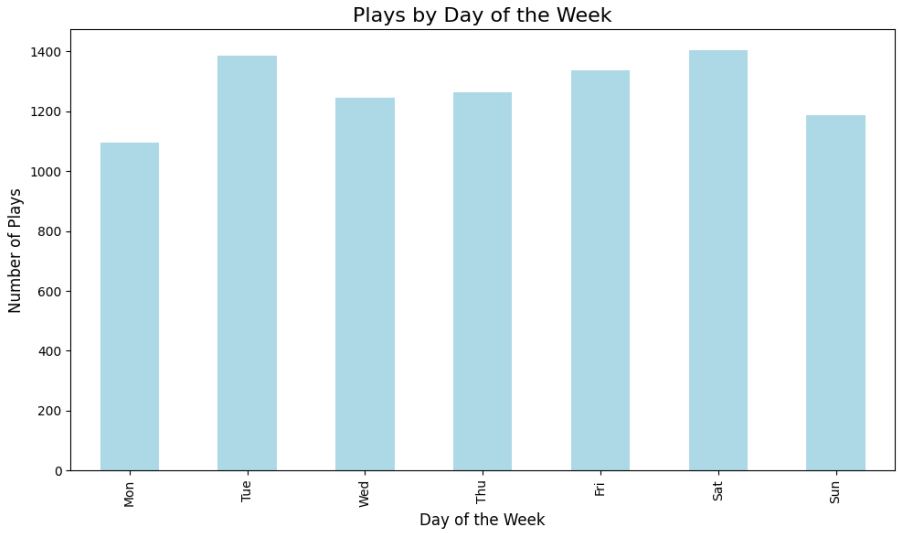

Figure 5: Plays by days of week

– **Monthly Listening Behavior**:

  * The highest listening activity was recorded in **June 2024**, with a total of 402 plays.
  * **November 2024** saw the least activity, attributed to academic or work commitments.

– **Reason**: The data can be explained by the following:
  * In June, I was working on a project in IIT Bhilai alone, hence the listening time suddenly spiked.
  * In October, my earphones broke, hence the sudden decline in the number of songs.
  * In November, my premium subscription ended, and the ads thereafter stopped me from listening to more music.
  * In December, I was back home, hence the listening time declined again.

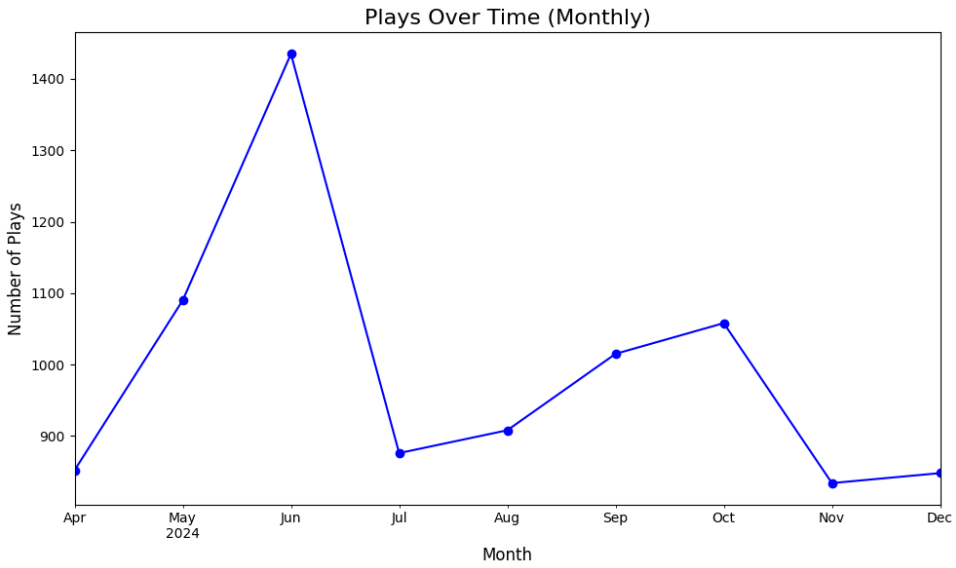

Figure 6: Plays Over Time(Monthly)

3.2. Hourly Patterns

– Listening activity was highest during the **evening hours (6 PM to 9 PM)**, correlating with leisure time after academic activites end at 5:30 PM.

– Late-night listening sessions were common on weekends, indicating a shift in routine and late sleeping times during the weekend.

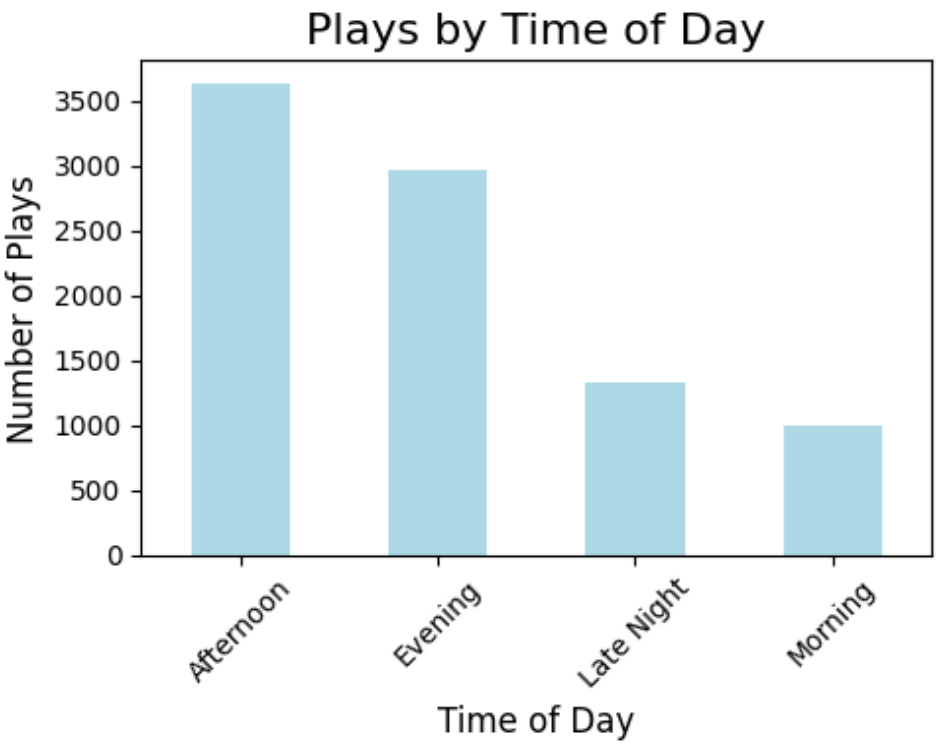

Figure 7: Plays by time of day

3.3. Peak Hours

- The top 3 peak hours were identified as 7 PM, 8 PM, and 9 PM.
- Visualized using **line plots**, with additional insights through radar charts showcasing hourly distributions.

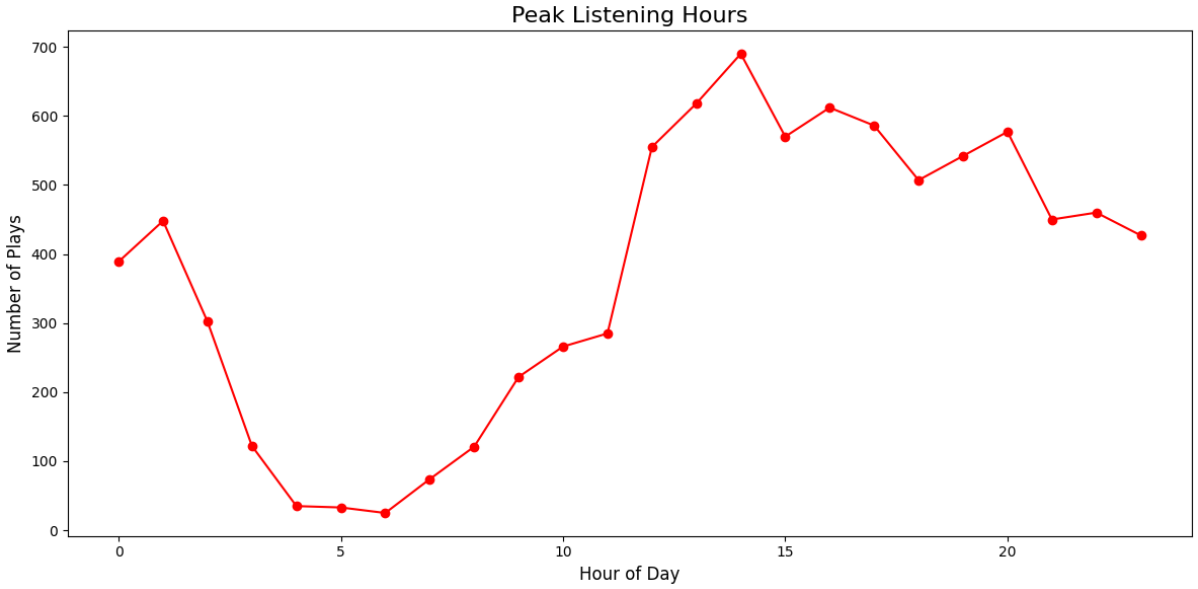

Figure 8: Peak Listening Hours

Insights:

- Listening habits align with typical patterns of relaxation, with evenings and weekends being the most active periods.
- Monthly peaks suggest specific events, seasons, or emotional states influence listening behavior.

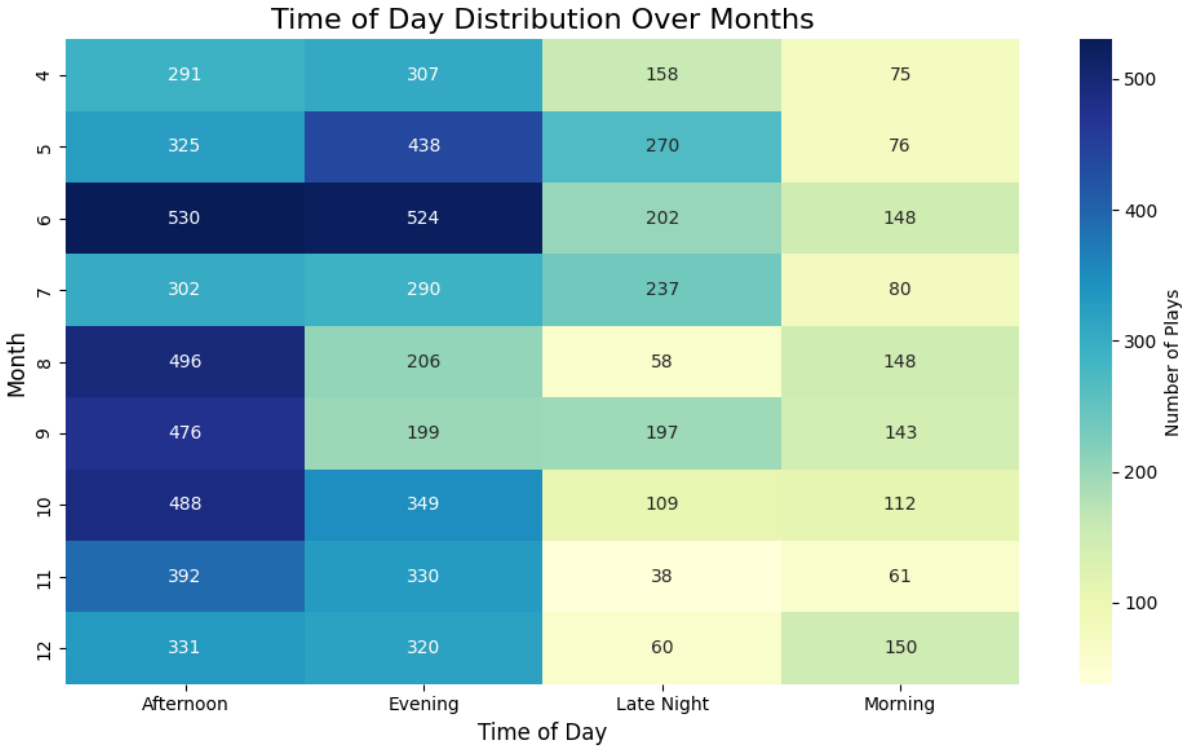

Figure 9: Time of Day Distribution Over Months

4. Favorite Artists, Songs, and Albums

4.1. Top Artists

- **Most Played Artists**:
  - **Anuv Jain**: 457 plays, with peak listening on June 4.
  - **Ariana Grande**: 399 plays, showcasing a broad appeal.
  - **Aditya Rikhari**: Consistent playtime across multiple months.
- Visualized using bar charts for the top 10 artists, revealing a preference for Indian and global indie artists.

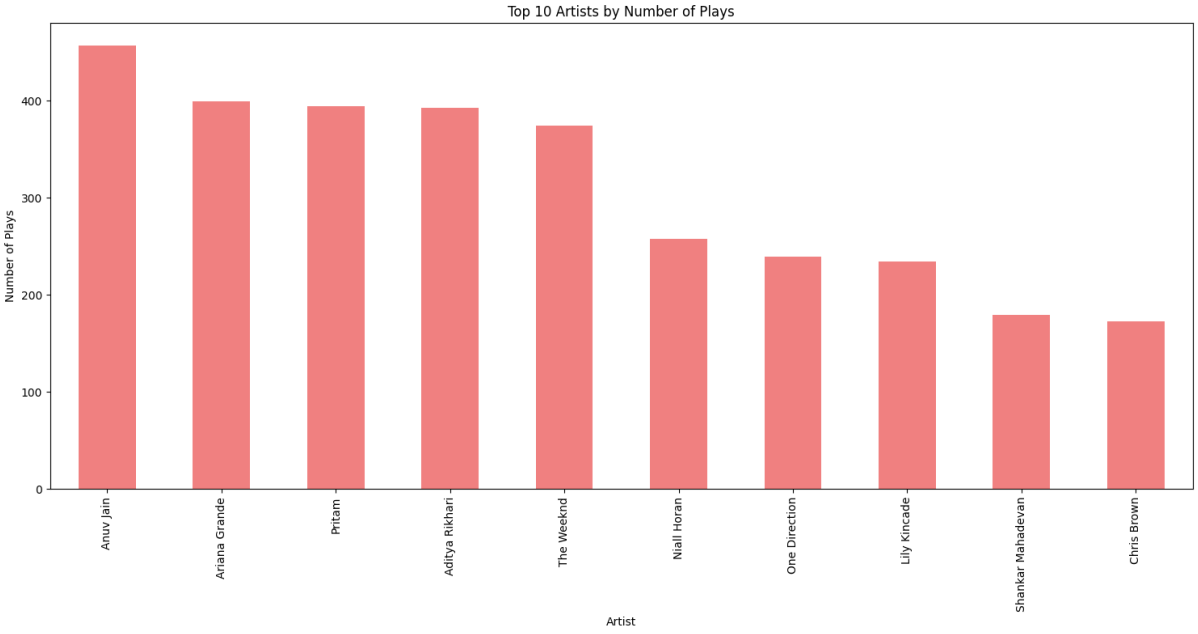

Figure 10: Top Artists

4.2. Top Songs

– **Consistent Favorites**:

* *Faasle* by Aditya Rikhari: Appeared in 4 months, averaging 63.5 plays.
* *Sweet n Low* by Lily Kincade: Appeared in 3 months, averaging 70.7 plays.
* *Daylight* by David Kushner: Appeared in 2 months, averaging 47.5 plays.
* *Co2* by Prateek Kuhad: Appeared in 2 months, averaging 35.5 plays.
* *Breathless* by Shankar Mahadevan: Appeared in 2 months, averaging 34.5 plays.
* *Samjho Na* by Aditya Rikhari: Appeared in 2 months, averaging 29.0 plays.

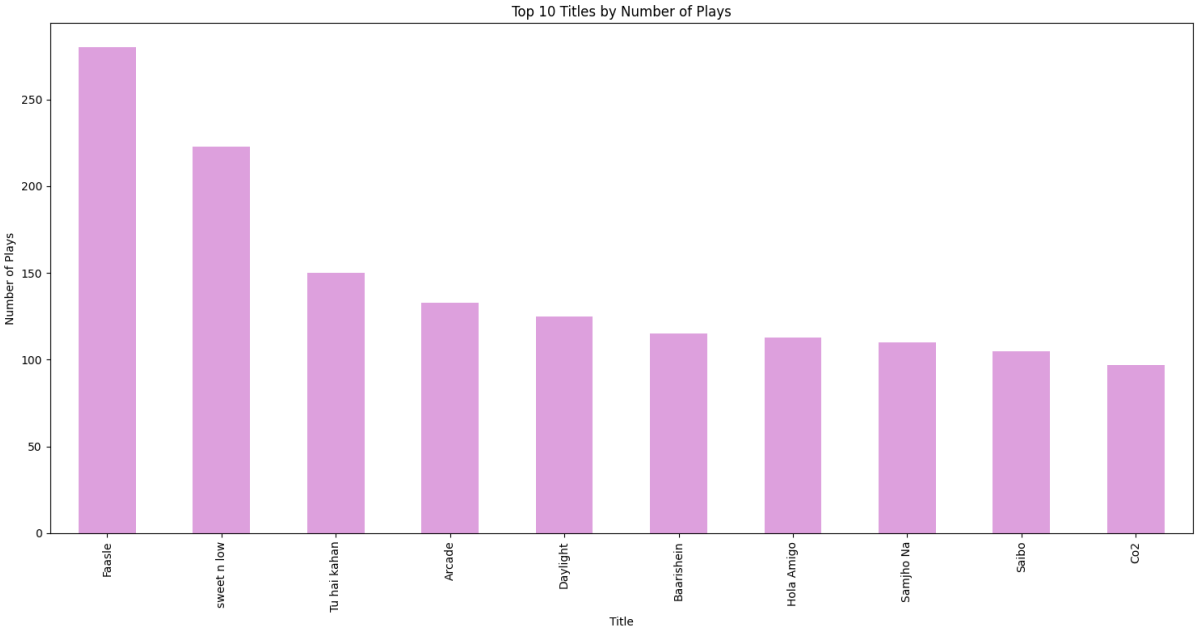

Figure 11: Top Songs

4.3. Top Albums

– Albums by Anuv Jain and Aditya Rikhari dominated the top 5, with significant daily plays.

4.4. Monthly Trends

– **2024-04**: 201 total plays, 5 unique artists, average plays per song: 40.2, top artist: Duncan Laurence.

– **2024-05**: 224 total plays, 5 unique artists, average plays per song: 44.8, top artist: Chris Brown.

– **2024-06**: 402 total plays, 4 unique artists, average plays per song: 80.4, top artist: Aditya Rikhari.

– **2024-07**: 99 total plays, 5 unique artists, average plays per song: 19.8, top artist: SZA.

– **2024-08**: 233 total plays, 5 unique artists, average plays per song: 46.6, top artist: Lily Kincade.

– **2024-09**: 269 total plays, 5 unique artists, average plays per song: 53.8, top artist: The Local Train.

– **2024-10**: 288 total plays, 4 unique artists, average plays per song: 57.6, top artist: Shankar Mahadevan.

– **2024-11**: 91 total plays, 4 unique artists, average plays per song: 18.2, top artist: Aditya Rikhari.

– **2024-12**: 74 total plays, 5 unique artists, average plays per song: 14.8, top artist: Lady Gaga.

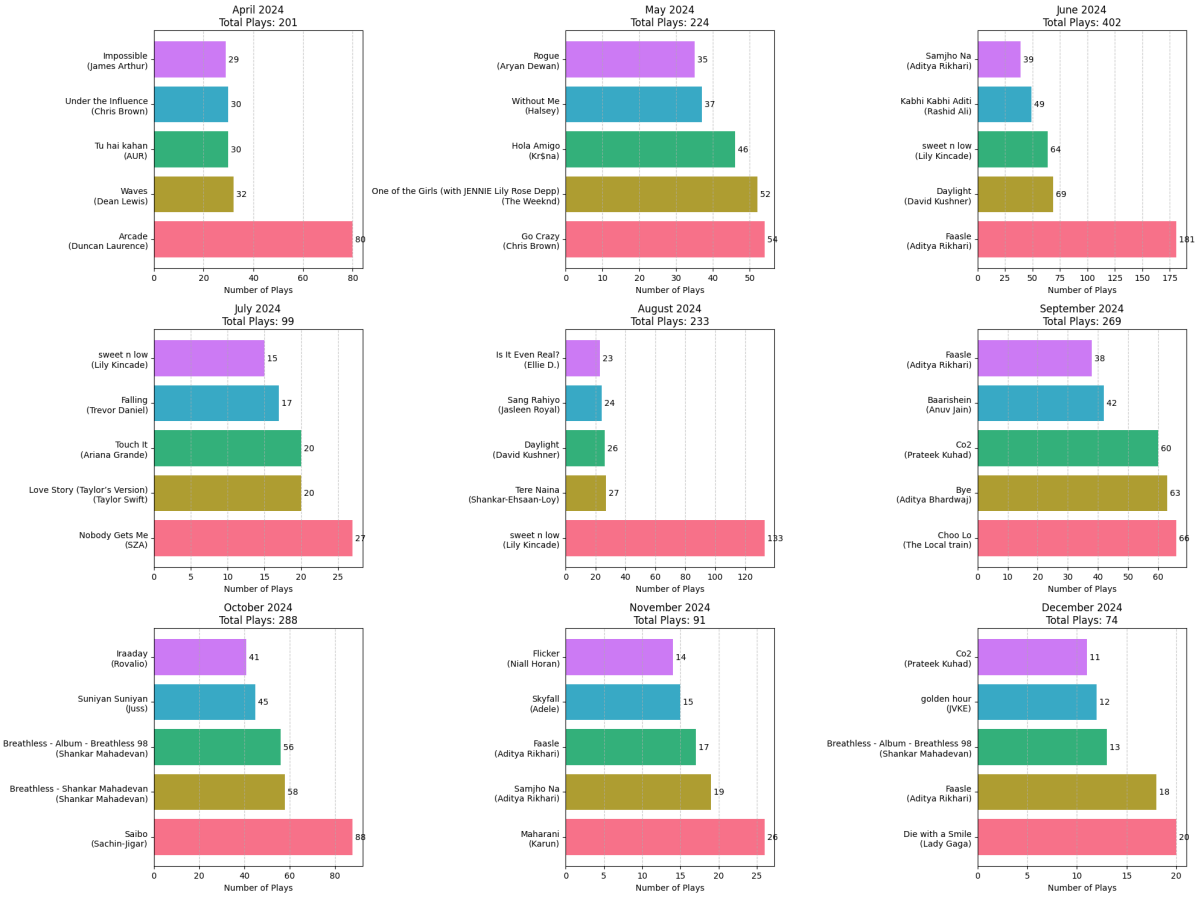

Figure 12: Monthly top songs

Insights:

- Repeated plays of specific songs and artists highlight emotional or thematic connections.
- A preference for a mix of Indian indie, Western pop, and alternative genres is evident.

5. Session Analysis

5.1. Definition of a Session

- **Criteria**:
  - A session was defined by a gap of fewer than 20 minutes between consecutive plays.
  - Minimum session size: 7 songs.

5.2. Metrics

- **Average Session Duration**: 62 minutes.

– **Average Songs per Session**: 15.35.
– **Most Active Period**: Afternoon and evening sessions on Tuesdays.

5.2.1. Session Analysis Summary

– **Total number of sessions**: 384
– **Average session duration**: 62.02 minutes
– **Average songs per session**: 15.35
– **Most common session start time**: Afternoon
– **Most active day**: Tuesday

5.3. Session Trends

– Visualized using boxplots and scatter plots.
– High artist diversity during longer sessions indicates exploratory listening behavior.

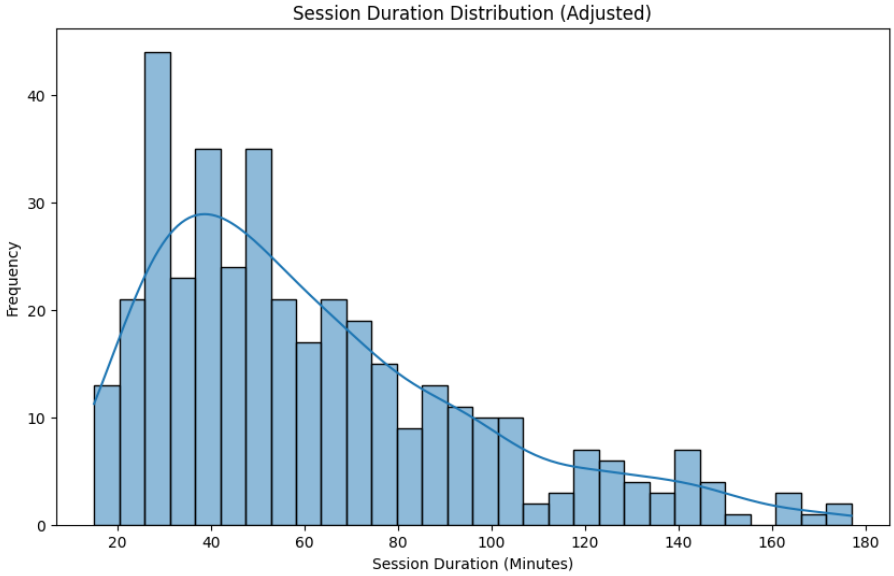

Figure 13: Session Duration Distribution

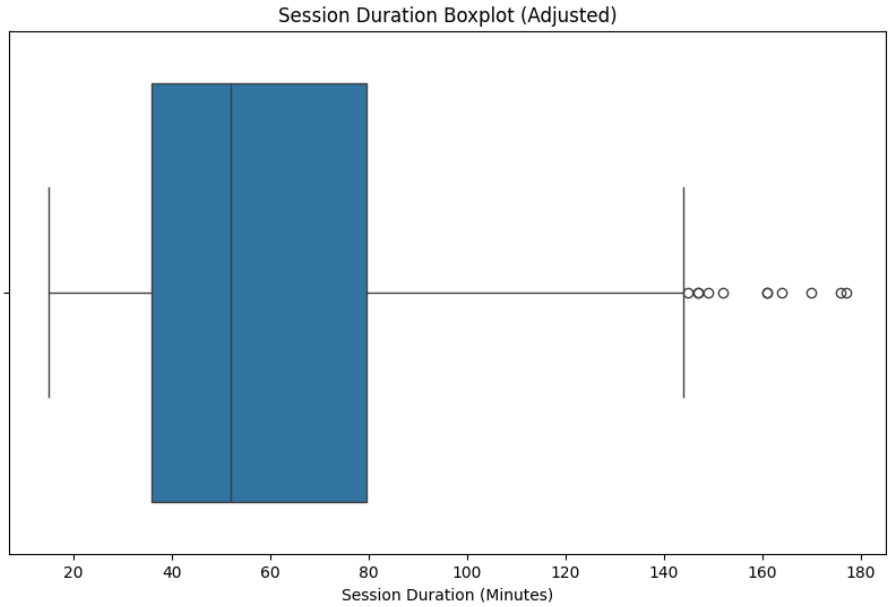

Figure 14: Session Box Plot

5.3.1. Engagement Metrics

– **Average songs per minute**: 0.26

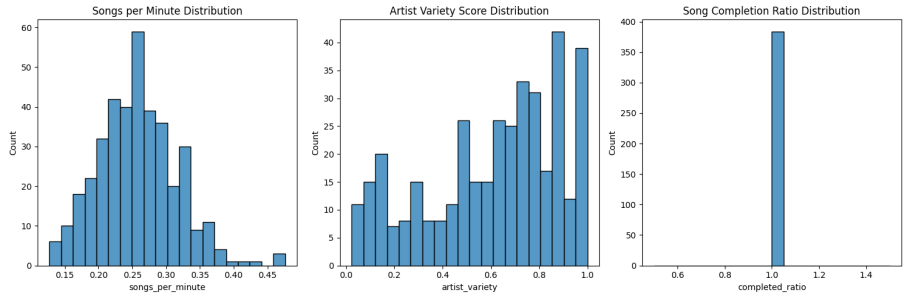

Figure 15: Engagement Metrics

Insights:

– Consistent afternoon sessions suggest music as a productivity or relaxation aid.

– Longer evening sessions align with leisure or social listening contexts.

– The high average number of songs per session and consistent start times (Afternoon, Tuesday) demonstrate that the listener engages in regular, extended listening sessions, often in structured time periods such as afternoons or evenings.

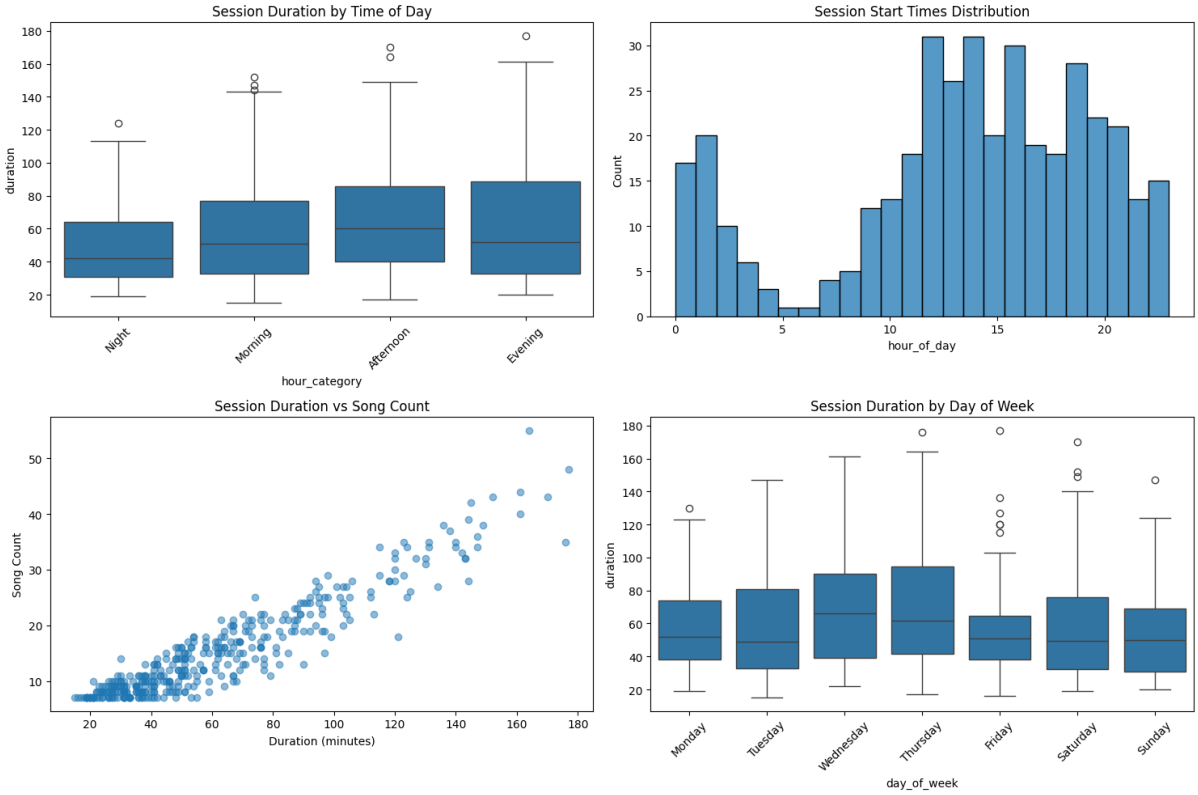

Figure 16: Enter Caption

6. Correlation Analysis

6.1. Artist Correlations

- **Strongest Positive Correlation**:
  - Anuv Jain and Anuv Jain (1.000): Perfect self-correlation.
- **Strongest Negative Correlation**:
  - Ariana Grande and Anuv Jain (-0.119): Contrasting listening patterns across different genres and moods.
- **Most Consistent Artist**:
  - Pritam, active on 124 days.
- **Highest Daily Play Intensity**:
  - Lily Kincade, with an average of 7.31 plays per active day.
- **Notable Patterns**:
  - Aditya Rikhari and Anuv Jain: 0.610
  - Ariana Grande and One Direction: 0.380

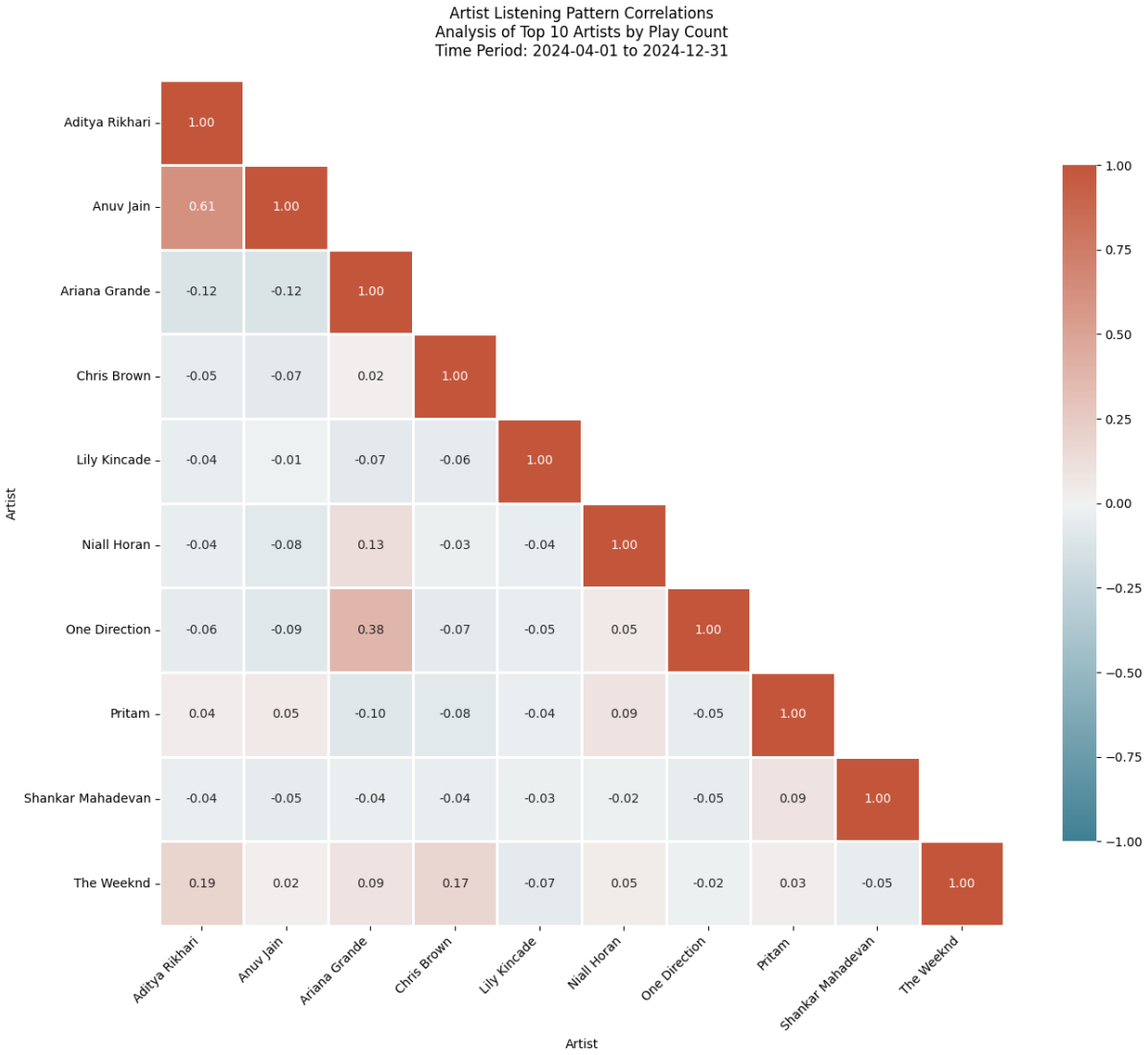

Figure 17: Artists Correlation

6.2. Song Correlations

- **Strongest Positive Correlation**:
  - 'Baarishein' by Anuv Jain with itself (1.000).
- **Strongest Negative Correlation**:
  - 'Arcade' by Duncan Laurence and 'Baarishein' by Anuv Jain (-0.128).
- **Most Consistently Played Song**:
  - 'Faasle' by Aditya Rikhari, active on 75 days.
- **Most Intensely Played Song**:
  - 'sweet n low' by Lily Kincade, with an average of 7.19 plays per active day.
- **Notable Patterns**:
  - 'Baarishein' (Anuv Jain) and 'Samjho Na' (Aditya Rikhari): 0.693

* 'Samjho Na' (Aditya Rikhari) and 'Baarishein' (Anuv Jain): 0.693

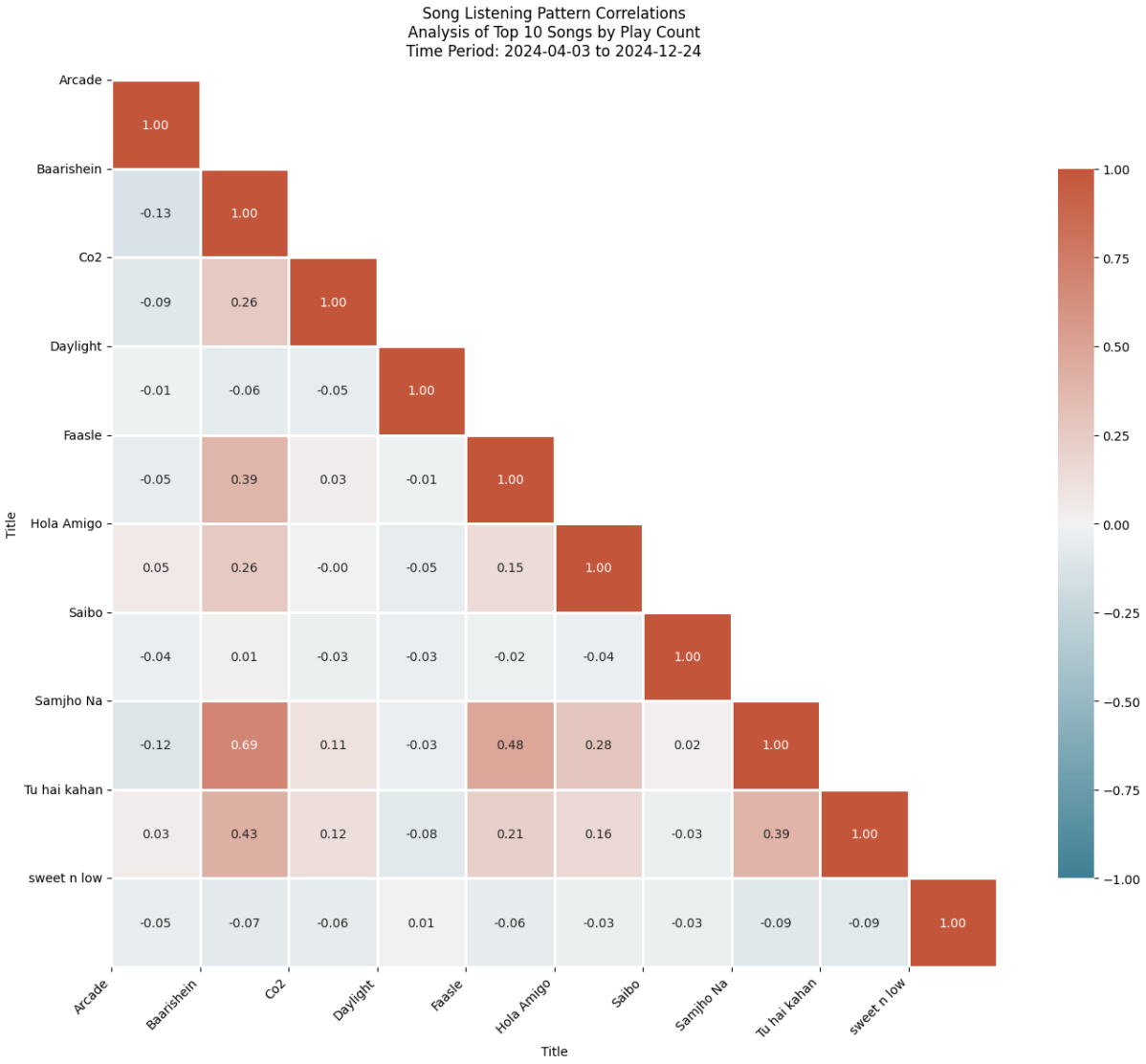

Figure 18: Songs Correlation

6.3. Song Transition Analysis

– **Strongest Song Transitions**:

* *HUSN* by Anuv Jain → *Tu hai kahan* by AUR (22.58%).
* *Baarishein* by Anuv Jain → *Tu hai kahan* by AUR (18.26%).
* *Baarishein* by Anuv Jain → *Faasle* by Aditya Rikhari (17.39%).

– **Transition Timing**:

* Average time between songs: 200.7 minutes.
* Median time between songs: 6.0 minutes.

– **Artist Transition Patterns**:

* Same artist transitions: 3.1%.

∗ Different artist transitions: 96.9%.

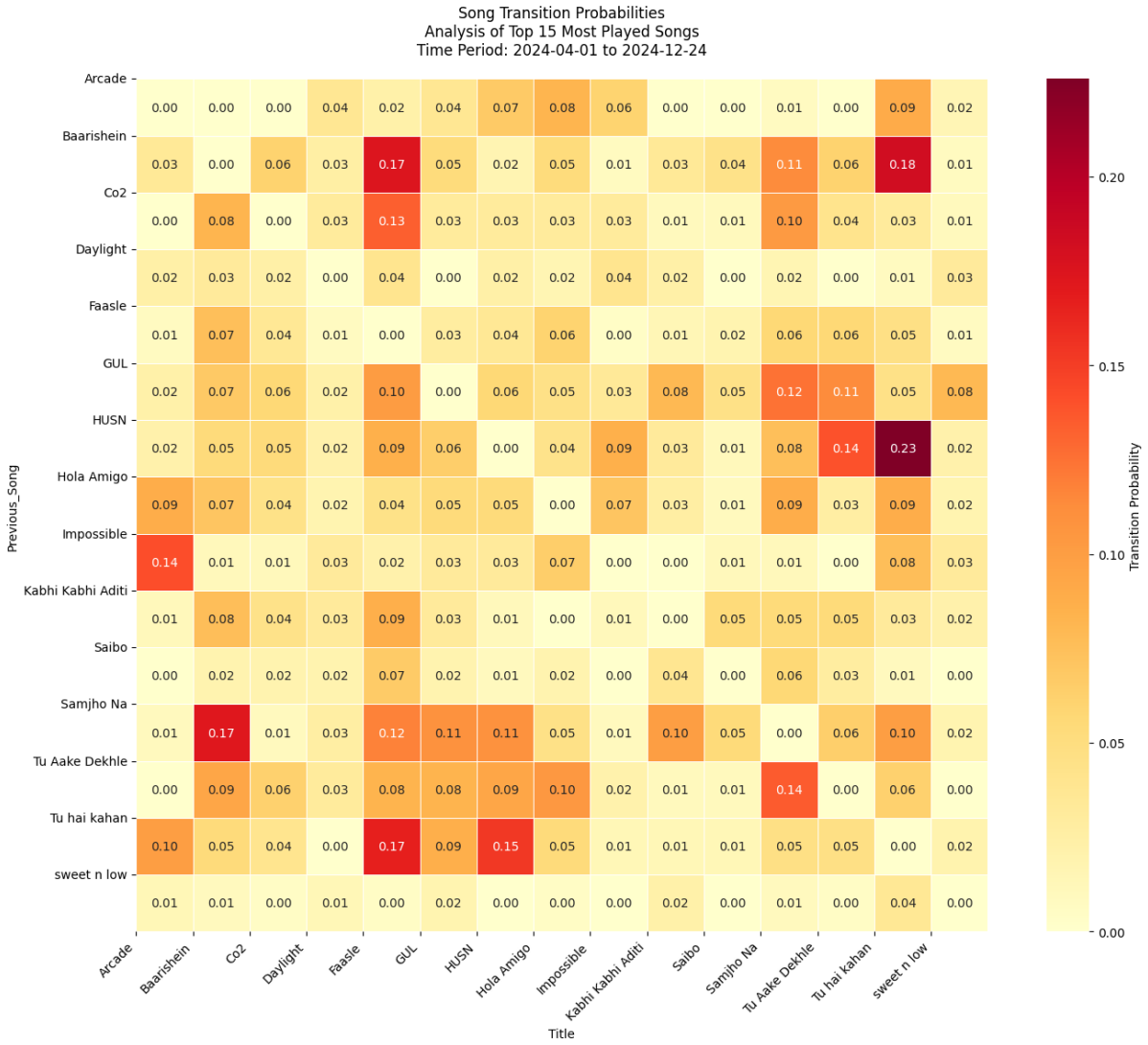

Figure 19: Song Transition Correlation

Insights:

- Strong transitions between emotionally resonant songs suggest mood-based playlists.
- Contrasting artist correlations point to varied listening contexts (e.g., focused vs. relaxing).
- High consistency of Pritam and Lily Kincade shows a preference for specific genres and mood states.

7. Discovery and Engagement Patterns

7.1. Discovery Rates

– A sharp rise in new artist discoveries during June, suggesting active exploration during this period.

7.1.1. Discovery Summary and Insights

– The listener actively explored new artists during June, with a significant increase in discoveries observed.

– This indicates a potential period of heightened interest in diverse music or genre exploration.

7.2. Engagement Metrics

– Repeat listens accounted for **87.6%** of all plays, indicating deep engagement with favorite tracks.

– The consistency score for overall listening was 0.31, highlighting regular habits.

7.2.1. Engagement Summary and Insights

– The high percentage of repeat listens (87.6%) reflects a strong preference for familiar songs, showing that the listener has a set of favorites that they consistently return to.

– The consistency score of 0.31 further confirms regular listening patterns, indicating moderate to high engagement over time.

7.3. Artist Loyalty

– **Metrics**:

  * Lily Kincade had the highest loyalty score (2.44), attributed to consistent daily plays.

7.3.1. Artist Loyalty Summary and Insights

– **Lily Kincade** leads with the highest loyalty score of 2.44, attributed to her frequent plays across a consistent period. This indicates the listener's deep connection to her music.

– Other artists like **Shankar Mahadevan**, **Niall Horan**, and **Juss** also exhibit strong loyalty, although at slightly lower levels compared to Lily Kincade.

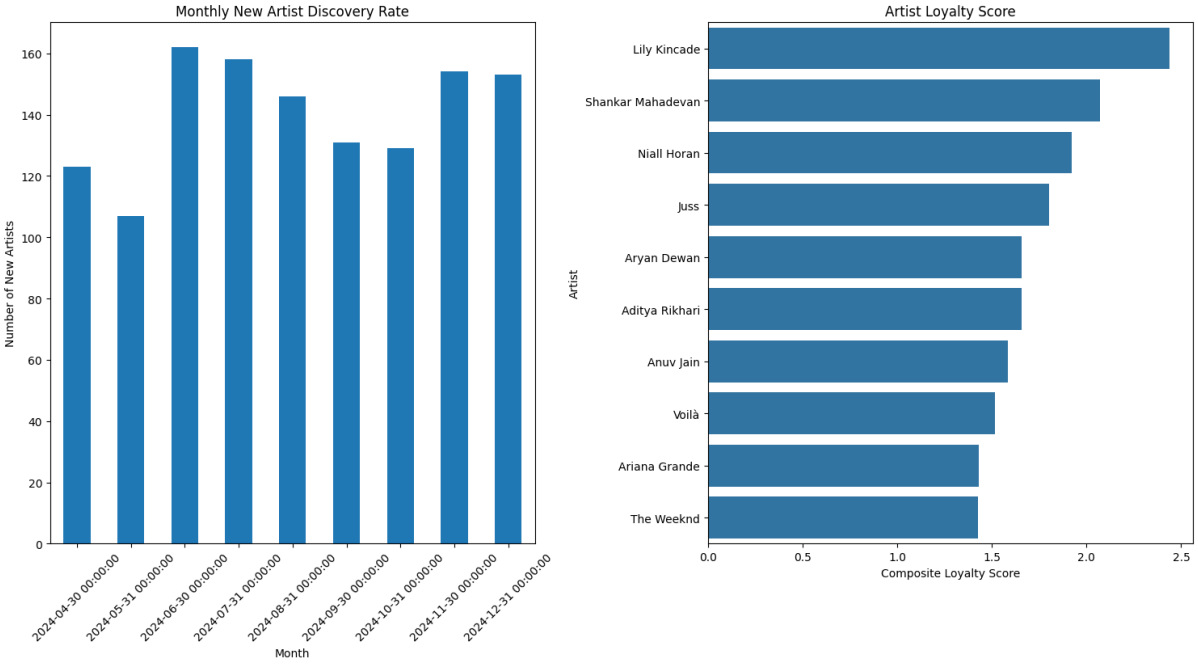

Figure 20: Artist Discovery Rate and Loyalty

7.4. Artist Trending Insights

- **Anuv Jain**: 457 total plays with an average of 1.66 plays per day. Peak on June 4.
- **Ariana Grande**: 399 total plays, averaging 1.45 plays per day. Peak on July 9.
- **Pritam**: 394 total plays, averaging 1.43 plays per day. Peak on September 24.
- **Aditya Rikhari**: 393 total plays, averaging 1.43 plays per day. Peak on June 3.
- **The Weeknd**: 374 total plays, averaging 1.36 plays per day. Peak on May 16.

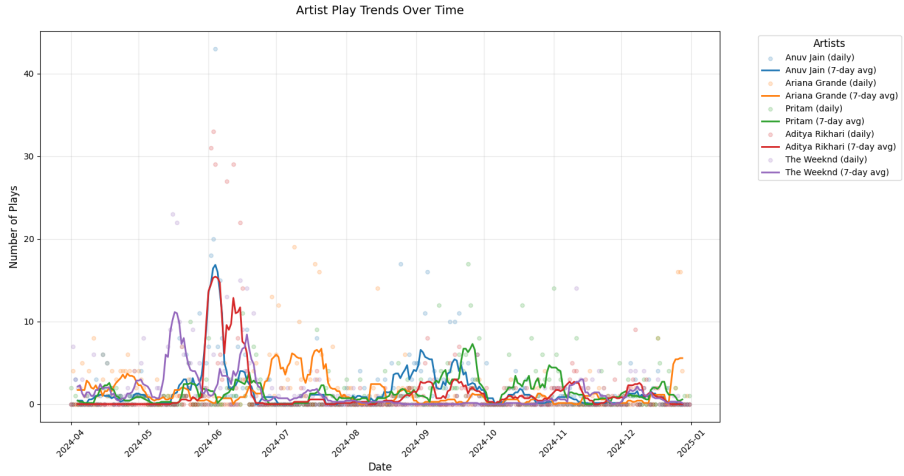

Figure 21: Artists play trend over time

7.5. Title Trending Insights

- *Faasle* by Aditya Rikhari: 280 plays with an average of 1.05 plays per day. Peak on June 3.
- *Sweet n Low* by Lily Kincade: 223 plays with an average of 0.84 plays per day. Peak on August 9.
- *Tu Hai Kahan*: 150 plays with an average of 0.56 plays per day. Peak on May 2.
- *Arcade*: 133 plays with an average of 0.50 plays per day. Peak on April 7.
- *Daylight*: 125 plays with an average of 0.47 plays per day. Peak on June 22.

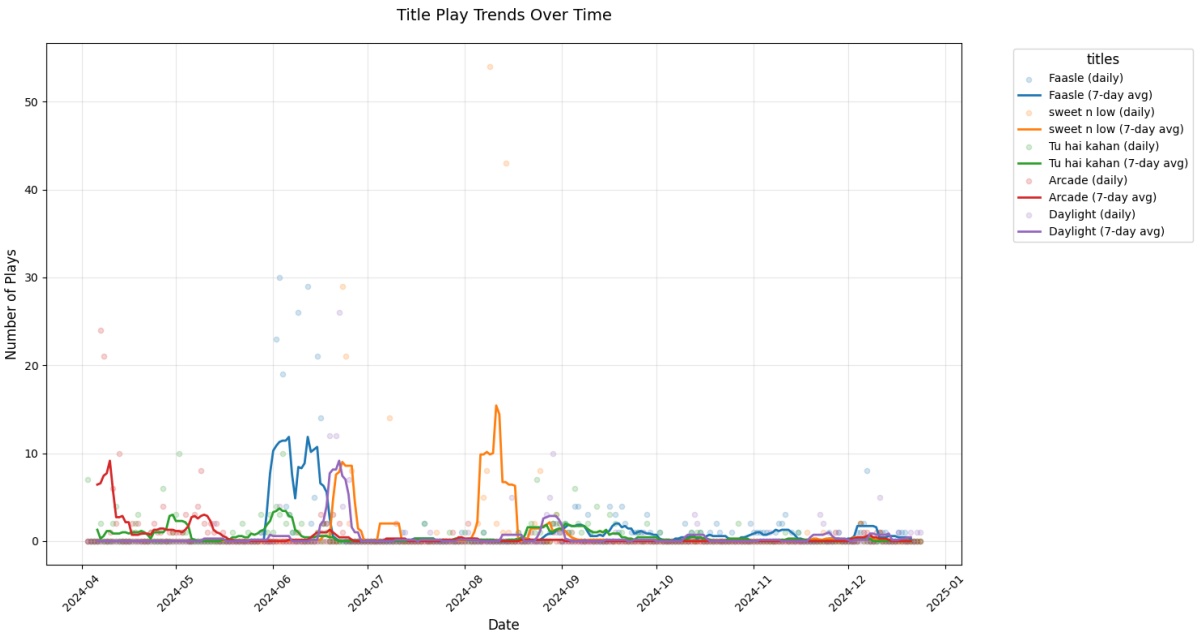

Figure 22: Title play trend over time

Insights:

- Discovery peaks align with changes in mood, seasons, or external influences.
- High repeat rates emphasize a preference for emotional familiarity.
- High loyalty scores for select artists indicate strong emotional connections.

8. Clustering Analysis

8.1. Hierarchical Clustering

- Artists were clustered based on hourly, daily, and monthly listening distributions, using hierarchical clustering with Ward's method.
- Distinct clusters emerged, separating indie, pop, and experimental artists, as well as artists with varying fanbase sizes and listening patterns.
- The hierarchical dendrogram provides insights into the relationships and similarity of artist listening patterns.

Cluster Analysis Results:

The clustering analysis identified the following groups:

Cluster	Size	Artists (Top 5)	Percentage of Total Artists
1	2	Niall Horan, One Direction	8%
2	1	David Kushner	4%
3	2	Rashid Ali, Sachin-Jigar	8%
4	10	Ariana Grande, Chris Brown, Duncan Laurence, H.E.R., Khalid	40%
5	10	AUR, Aditya Rikhari, Anuv Jain, King, Lily King	40%

Table 1: Clustering Results and Artist Distribution

Insights:

- The clustering highlights the diversity in artist listening patterns, with larger clusters corresponding to well-known pop and mainstream artists, while smaller clusters feature indie and experimental genres.
- The larger clusters (Clusters 4 and 5) represent a significant portion (80%) of the artists, indicating strong groupings based on genre or popularity.
- Smaller clusters (Clusters 1, 2, and 3) suggest more niche or unique listening patterns, likely reflecting distinct fanbase preferences and music styles.

Dendrogram Visualizations:

- The complete hierarchical dendrogram and truncated dendrogram provide a clear visual representation of the clusters. The full dendrogram showcases the distance between clusters, while the truncated version highlights the most distinct clusters.

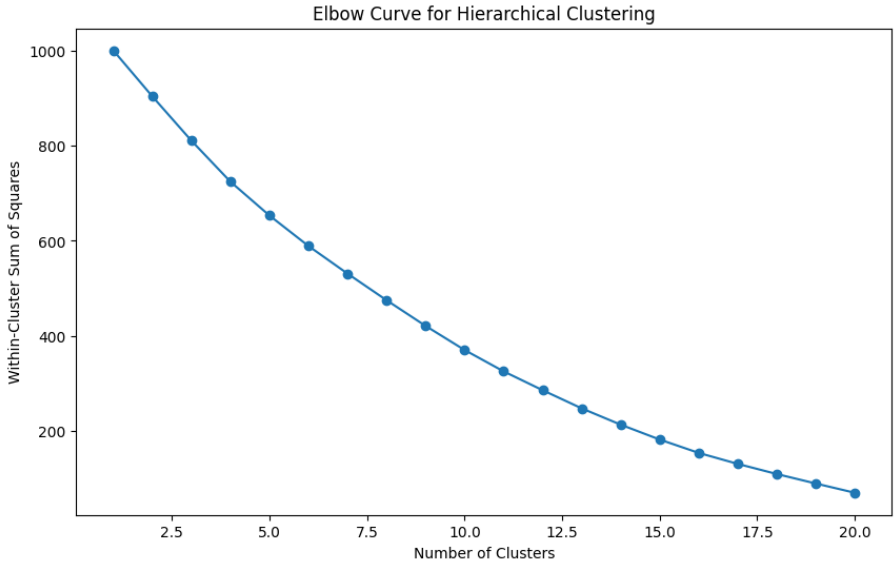

Figure 23: Cluster Elbow Curve

– An elbow curve analysis was performed to determine the optimal number of clusters. The elbow suggests a well-defined break in the data around 5 clusters, which was confirmed by further cluster analysis.

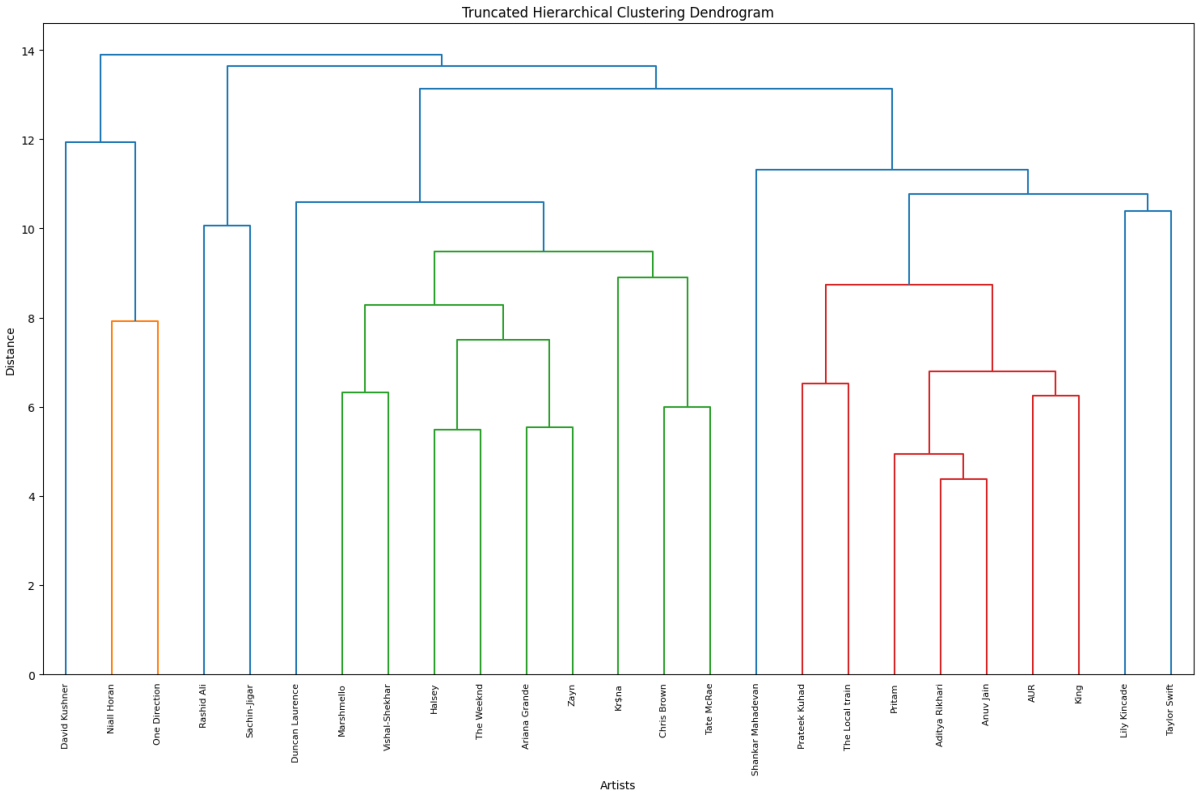

Figure 24: Cluters Dendogram

9. Conclusion

This analysis of Spotify listening data reveals intricate patterns in music consumption, driven by emotional, temporal, and contextual factors. Insights into artist loyalty, session behavior, and transition dynamics offer a deeper understanding of personal listening habits. Future studies could incorporate genre-specific analysis and predictive models to further enhance understanding.