SEGMENTATION FOR A SUBSCRIPTION-BASED STREAMING SERVICE

AGENDA

01. Business Problem

02. Problem Formulation

03. Collect and Label Data

04. Evaluate Data

05. Feature Engineering

06. Select and Train Model

07. Tune Model

08. Evaluate Model

09. Business Goal Check

Business Problem

A subscription-based streaming service wants to group users based on their listening behaviour and activity patterns.

Purpose

- Design personalised marketing campaigns
- Improve user retention and engagement

PROBLEM FORMULATION

Machine Learning Task: **Unsupervised Learning**

Technique: **Clustering / Segmentation**

No predefined target labels

Collect and Label Data

Dataset: Spotify Global Streaming Data (2024)

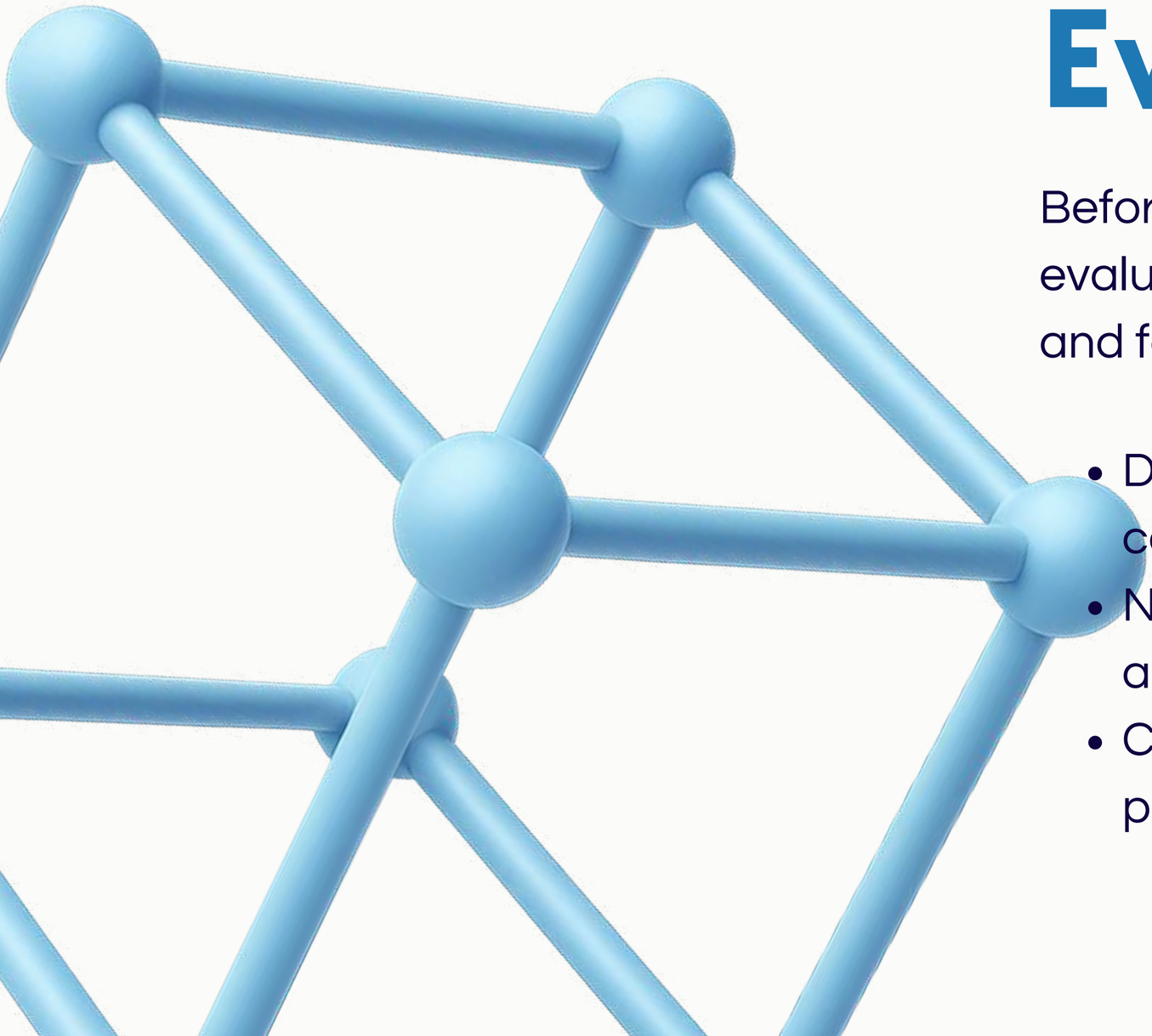
Source: Kaggle

Data type: Streaming behaviour and engagement metrics

Labels

- No labels are created
- Segmentation is performed without predefined classes

The dataset used in this project represents real-world streaming activity from a global music platform.



Evaluate Data

Before analysis and modelling, the dataset is evaluated to understand its size, structure, and feature types.

- Dataset contains both numerical and categorical features
- Numerical features describe listening activity and engagement
- Categorical features describe content and platform information

Data Evaluation and Quality Checks

Dataset Size

- **Dataset Shape Check**

We first checked the dataset shape by identifying the total number of rows and columns. This helps us understand how much data is available and whether the dataset is suitable for analysis.

Feature Types

- **Feature Type Identification**

Next, we identified and separated numerical features, such as streams and listeners, from categorical features like genre and country. This is important because each type of data is processed differently.

Missing Values

- **Missing Value Check**

We then checked the dataset for missing values to ensure that all records are complete. Missing data can affect analysis accuracy and lead to unreliable results if not handled properly.

Duplicates

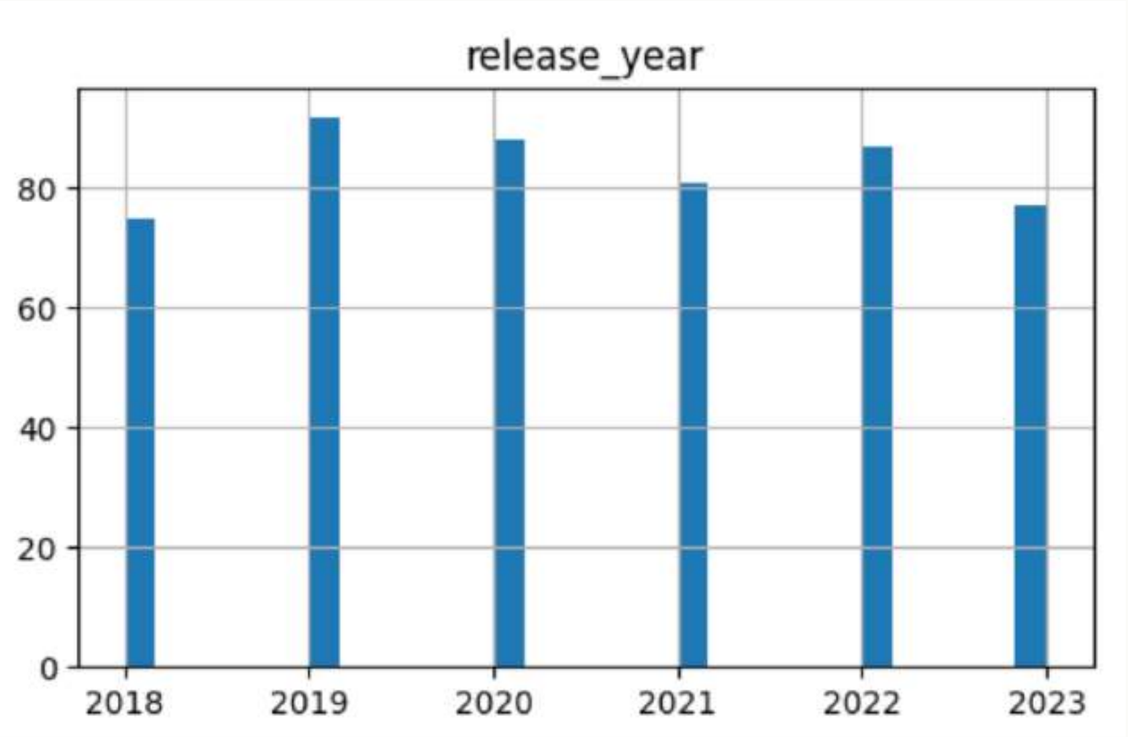
- **Duplicate Record Check**

After that, we checked for duplicate records in the dataset. Removing duplicates prevents repeated data from biasing patterns and misleading the analysis.

Value Consistency

- **Value Range Verification**

Finally, we verified that all values fall within reasonable and realistic ranges. For example, release years and skip rates were checked to ensure they make sense in a real-world context.



Release Year Distribution

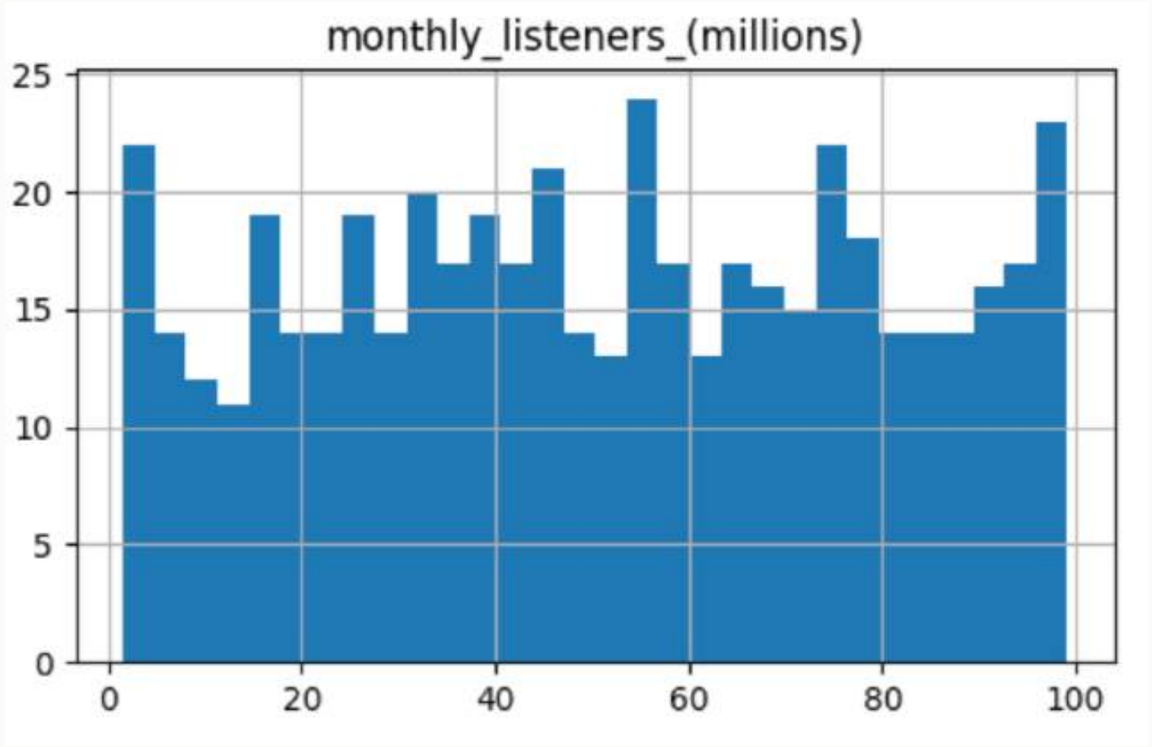
This chart shows how the songs in the dataset are distributed based on their release year.

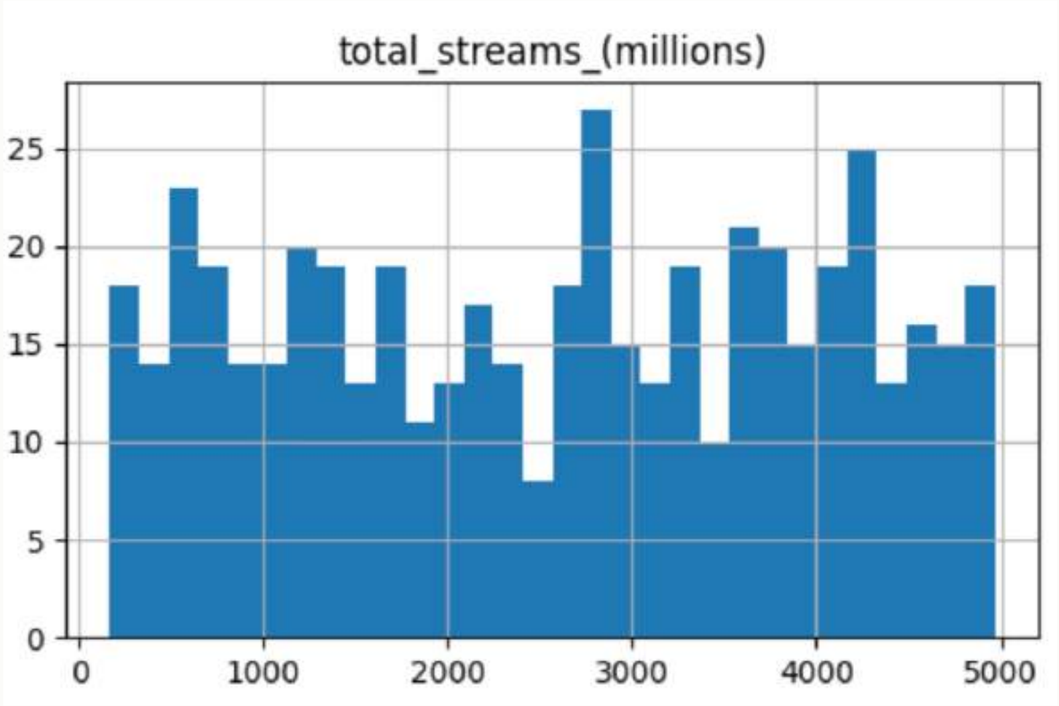
- Release years range from 2018 to 2023
- Focus on recent music
- Reflects modern streaming behaviour

Monthly Listeners Distribution

This chart shows how many monthly listeners different songs attract on the platform.

- Values range from close to 0 up to around 100 million
- Distribution is widely spread
- Indicates strong variation in popularity

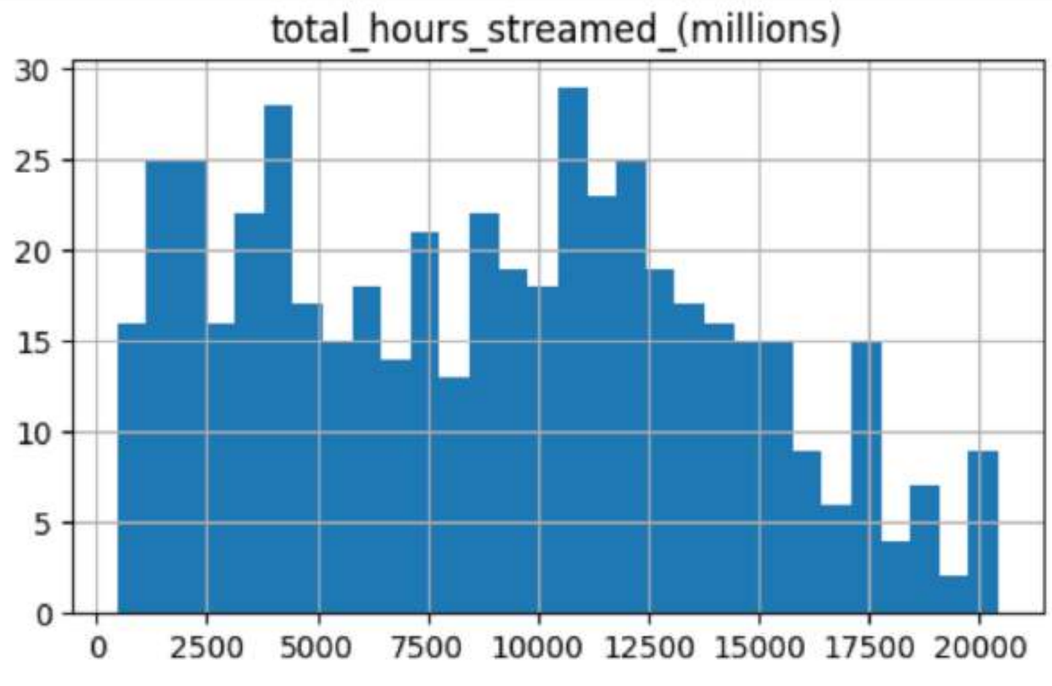




Total Streams Distribution

This chart shows how often songs have been streamed overall on the platform.

- Values range from near 0 up to around 5,000 million
- Distribution is highly spread
- Represents long-term popularity



Total Hours Streamed Distribution

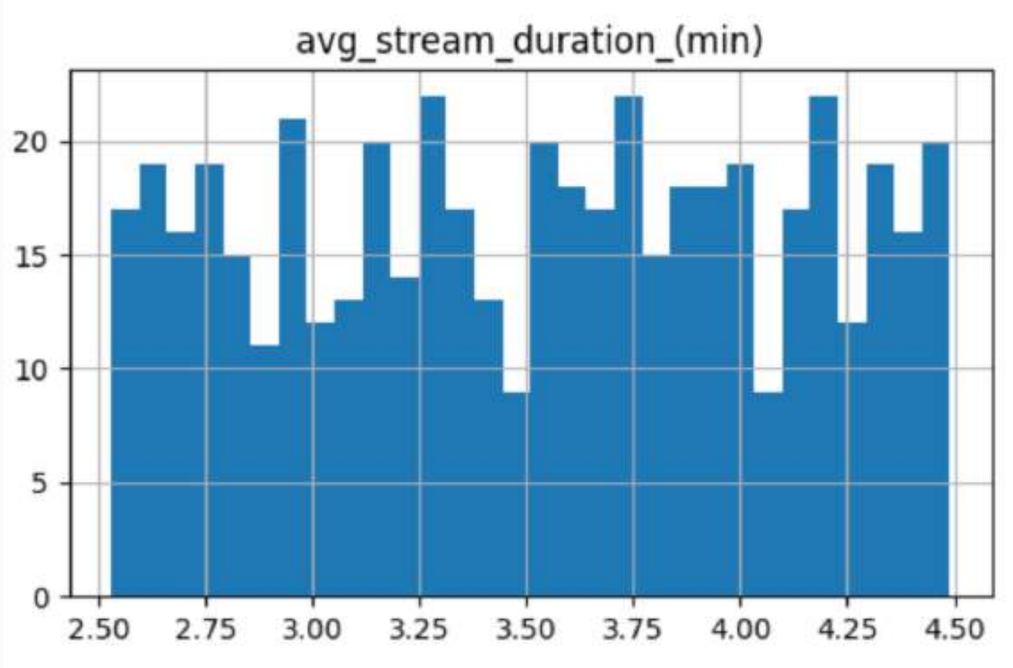
This chart shows the total amount of time users spend listening to songs on the platform.

- Values range from near 0 up to around 20,000 million hours
- Very wide distribution
- Reflects listening depth, not just play count

Average Stream Duration

This chart shows the average time users spend listening to a song per stream.

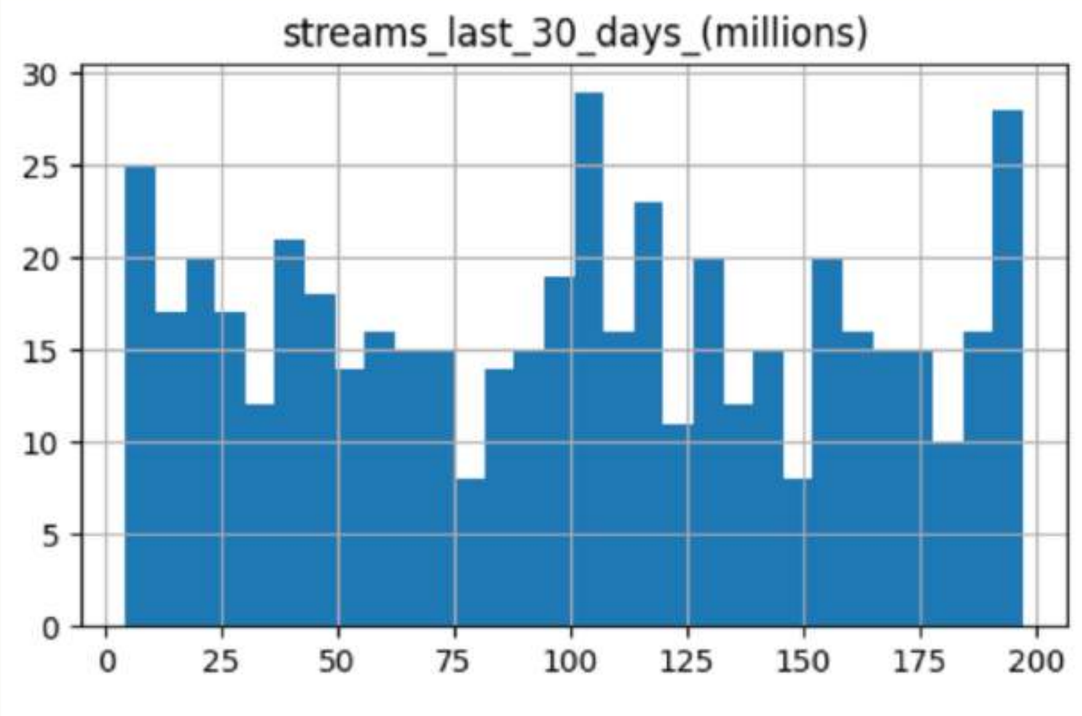
- Values range from about 2.5 to 4.5 minutes
- Narrower distribution compared to other features
- Indicates relatively consistent listening behaviour

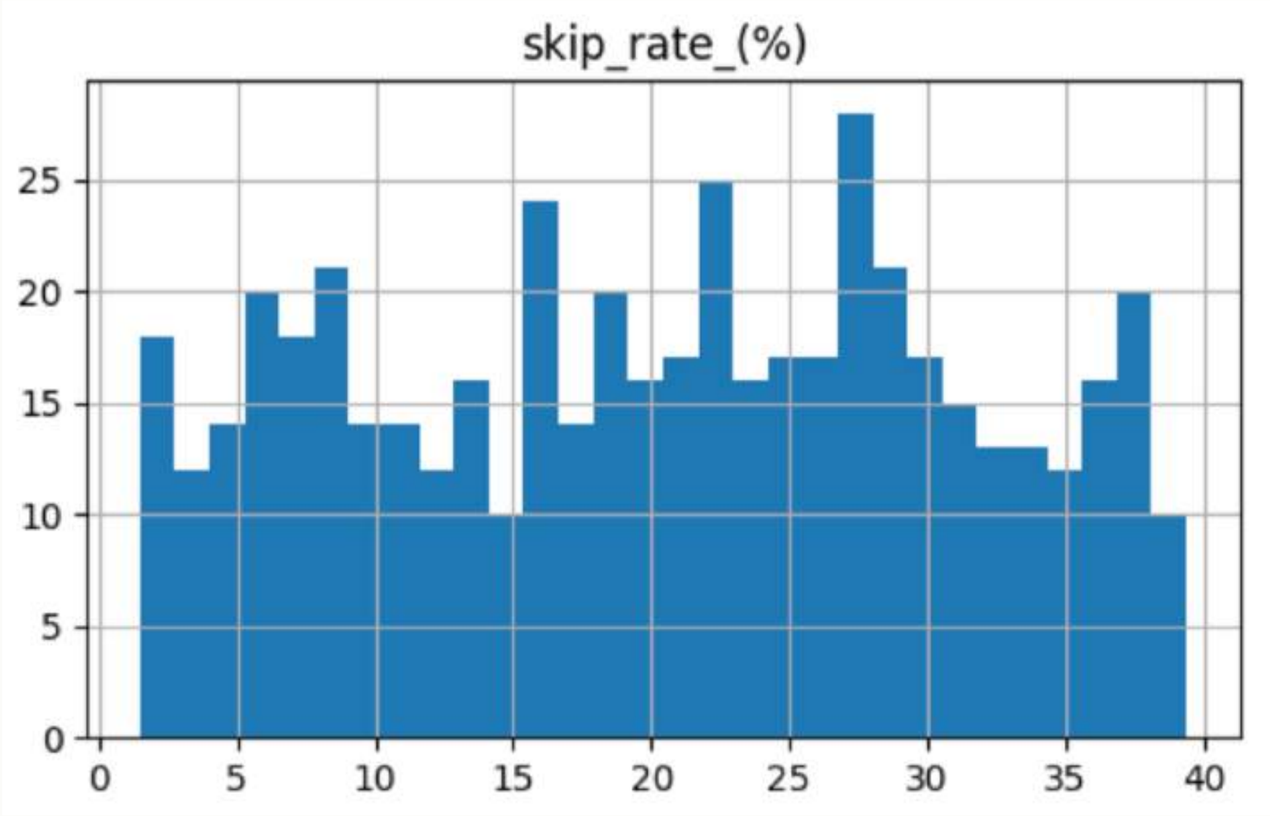


Streams in the Last 30 Days

This chart shows how often songs have been streamed recently.

- Values range from near 0 up to around 200 million
- Represents recent listening activity
- Highlights currently trending content

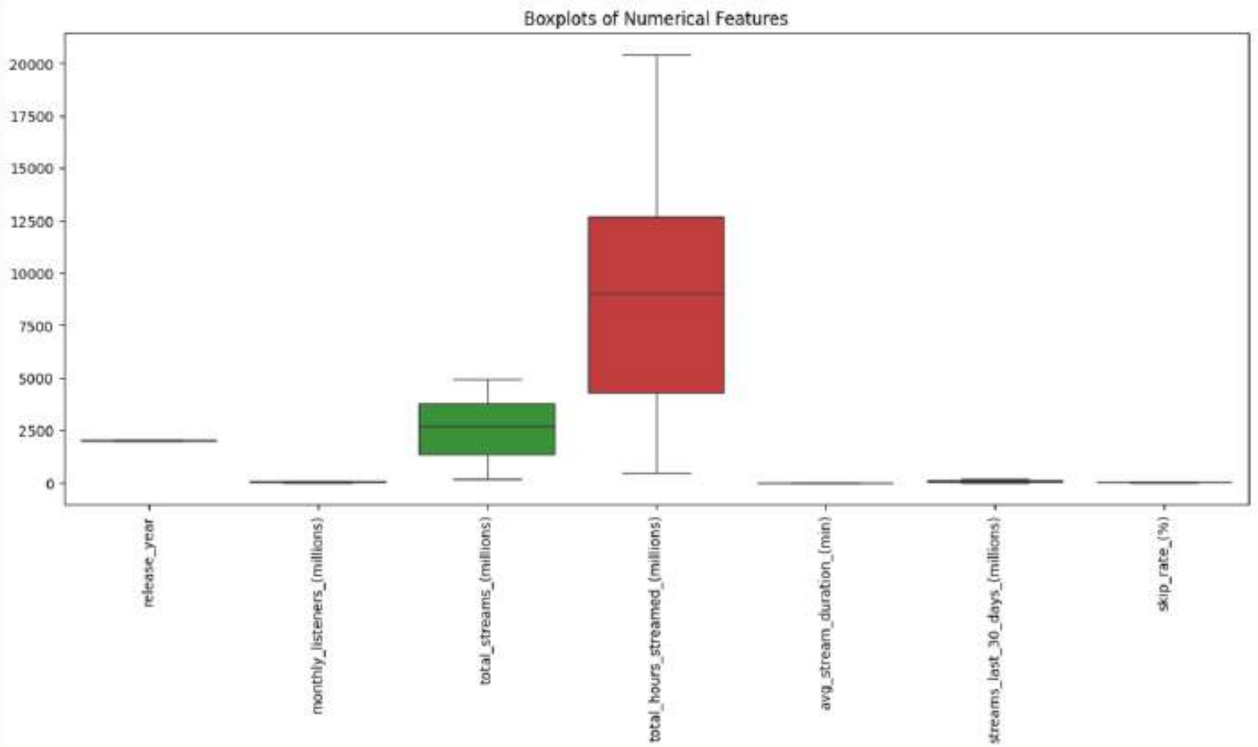




Skip Rate Distribution

This chart shows how often users skip songs before finishing them.

- Values range from about 0% to 40%
- Wide variation across songs
- Indicates differences in engagement quality

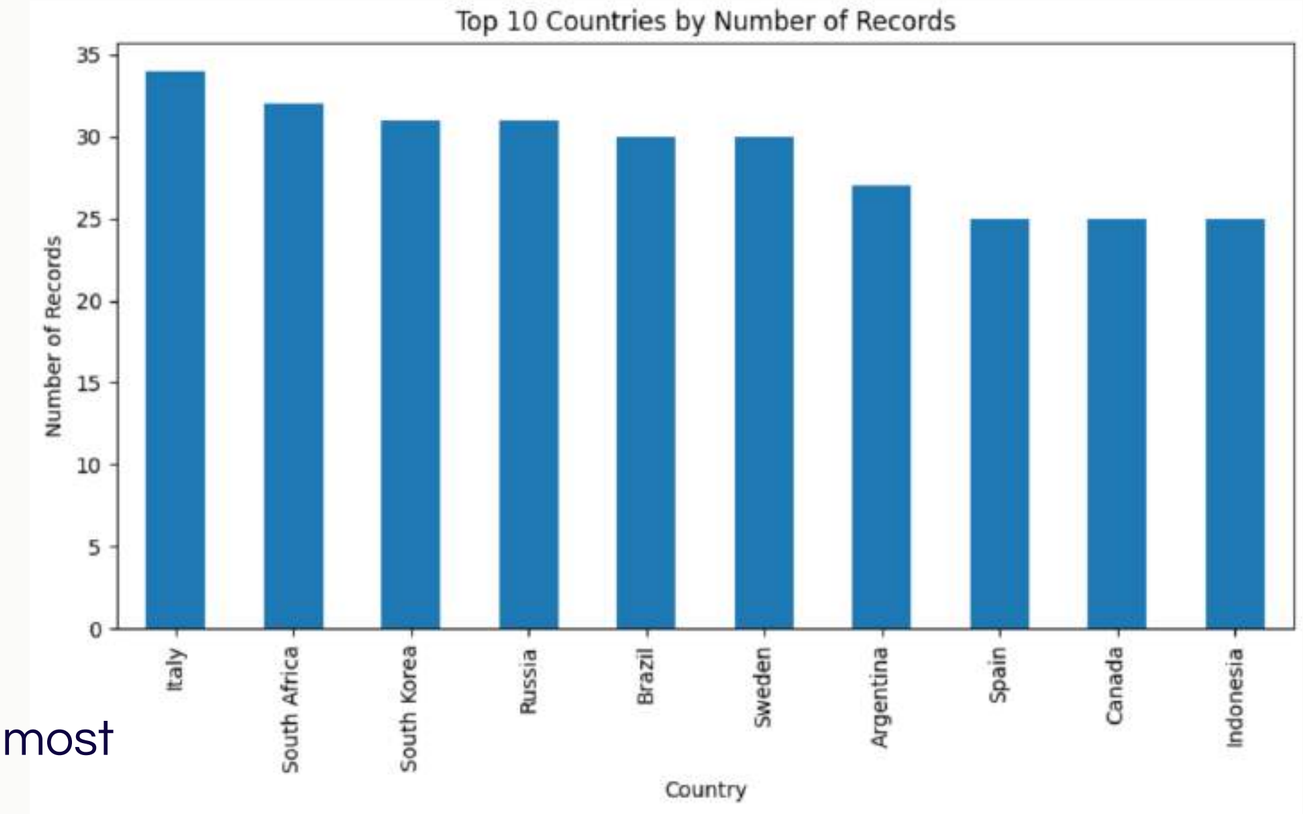


Boxplot Analysis of Numerical Features

This chart compares the spread and variability of all numerical features in the dataset.

- Streams and hours show the widest spread
- Some features contain extreme values
- Confirms differences in scale across metrics

Country Distribution



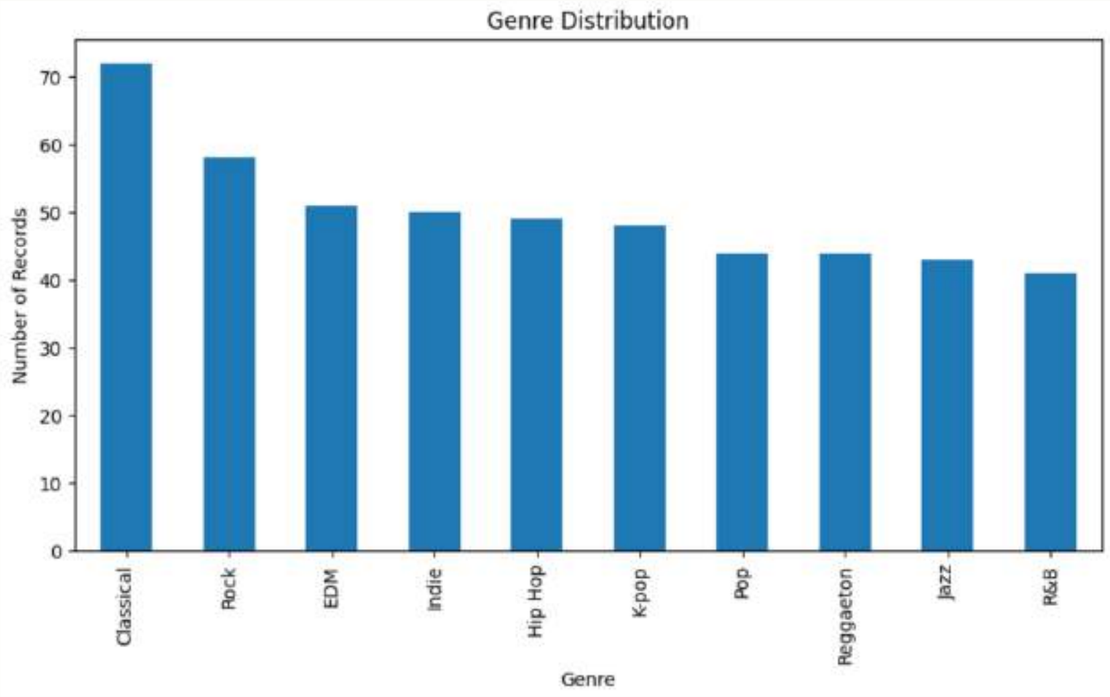
This chart shows the countries that appear most frequently in the dataset.

- Italy has the highest count (around 34 records)
- Other countries range between about 25 and 32 records
- Dataset represents multiple regions

Genre Distribution

This chart shows how songs are distributed across different music genres.

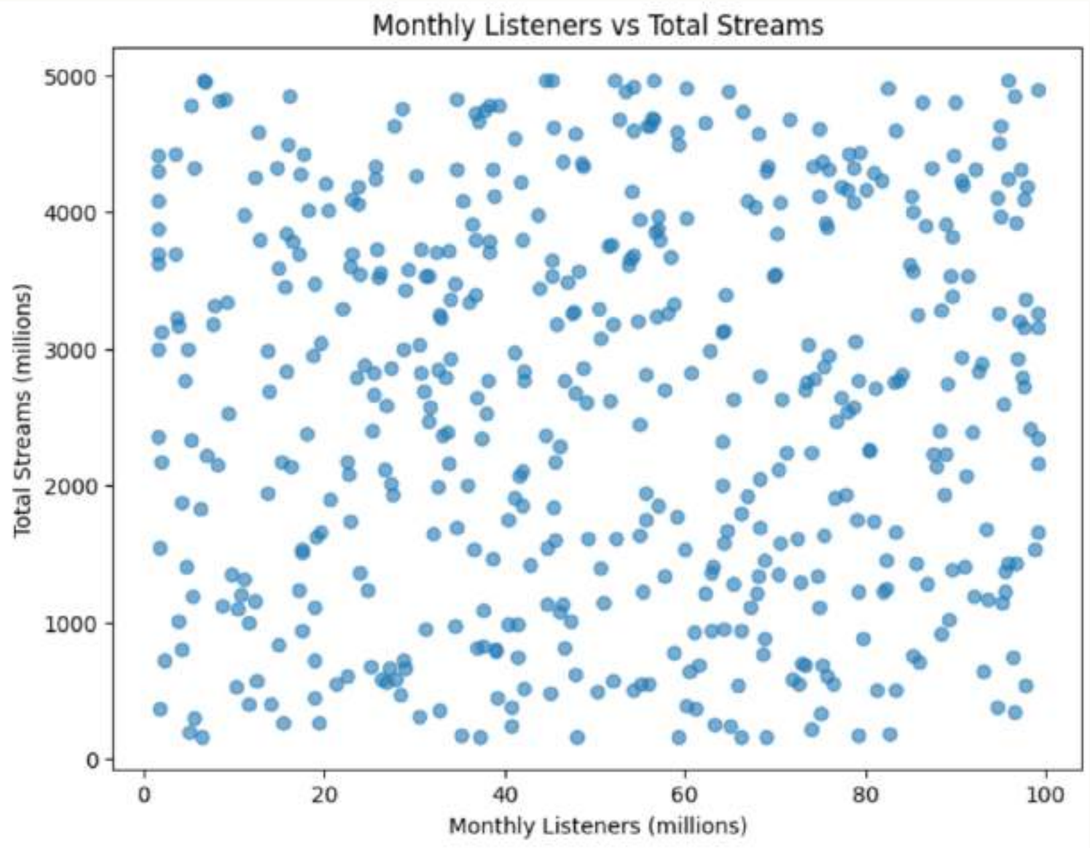
- Classical has the highest count (around 72 records)
- Rock follows with about 58 records
- Other genres range between around 40 and 51 records



Monthly Listeners vs Total Streams

This chart shows the relationship between audience size and overall popularity.

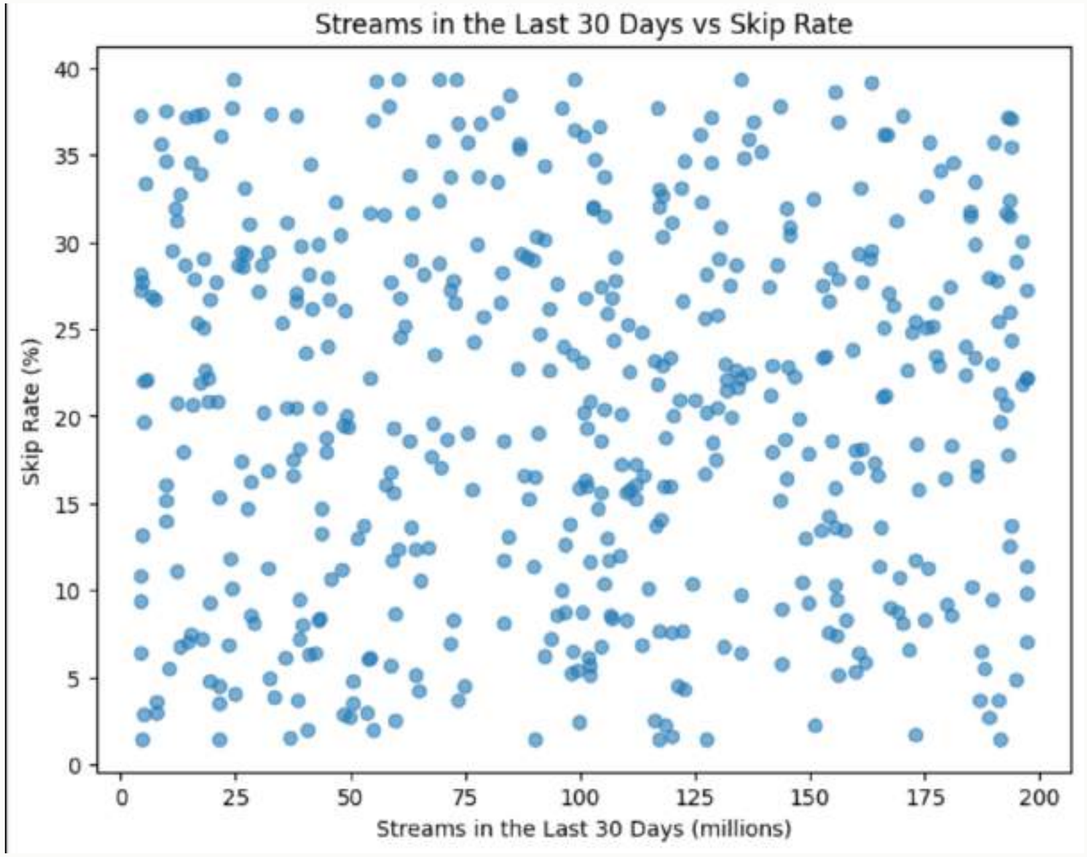
- General upward trend
- Wide spread of points
- Indicates different engagement patterns



Streams in the Last 30 Days vs Skip Rate

This chart compares recent popularity with how often users skip songs.

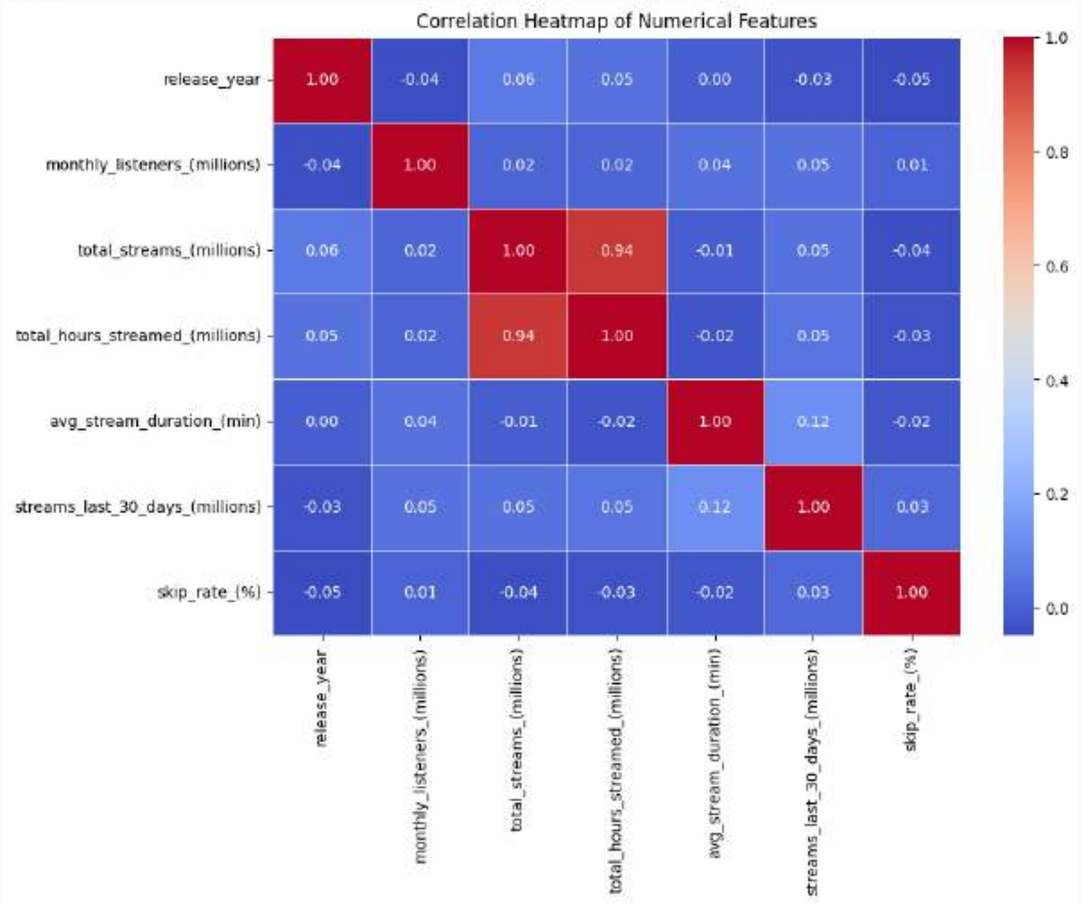
- No strong linear relationship
- Skip rate varies across all popularity levels
- Popularity and engagement quality differ



Correlation Analysis

This chart shows how strongly numerical features are related to each other.

- Strong correlation between Total Streams and Total Hours Streamed
- Most other feature relationships are weak
- Features capture different aspects of behaviour



Feature Engineering

Based on the data evaluation and exploratory analysis, new features are created to better represent user behaviour.

- Raw features capture different behaviour patterns
- Some features overlap, others provide unique insights
- Feature engineering improves segmentation quality

Engineered Features

To support effective segmentation, the following features were created or transformed:

- Genre encoded into numerical form
- Platform type converted to subscription category
- Engagement-based features derived from streams and listeners
- Recency-based feature created from recent activity

Select & Train Model

Models Used

- K-Means Clustering
- Hierarchical Clustering (Ward linkage)

Features Included

- "genre_code"
- "is_premium"
- "hours_per_listener"
- "recent_stream_ratio"

Standardize The Features

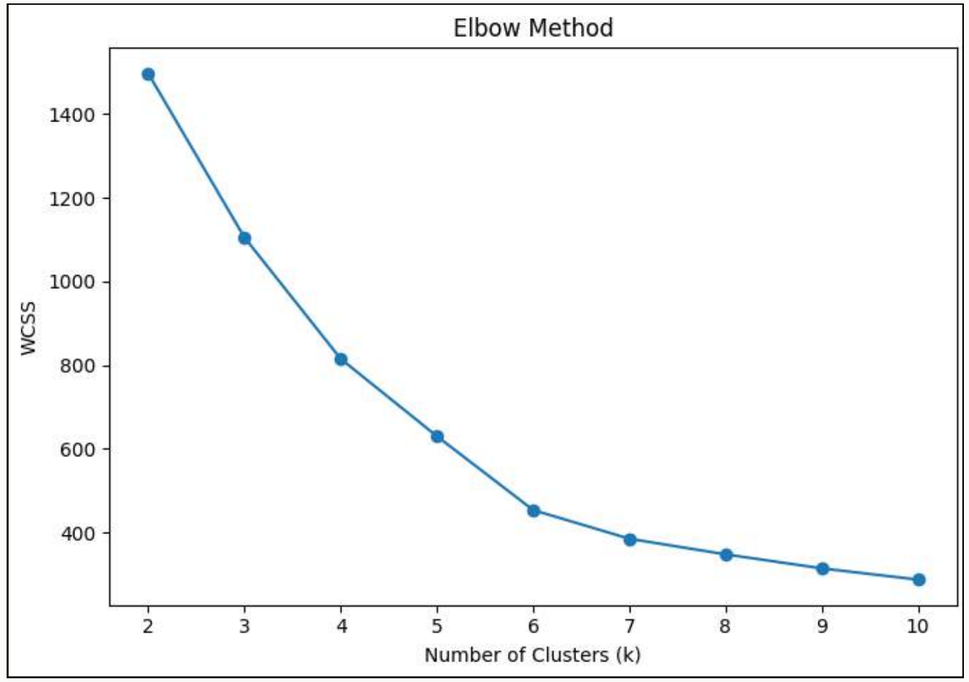
Feature means vary widely across variables

- After standardization, the scaled feature means are approximately 0, which is expected due to floating-point precision.
- The scaled feature standard deviations are approximately 1 for all features, indicating successful normalization.

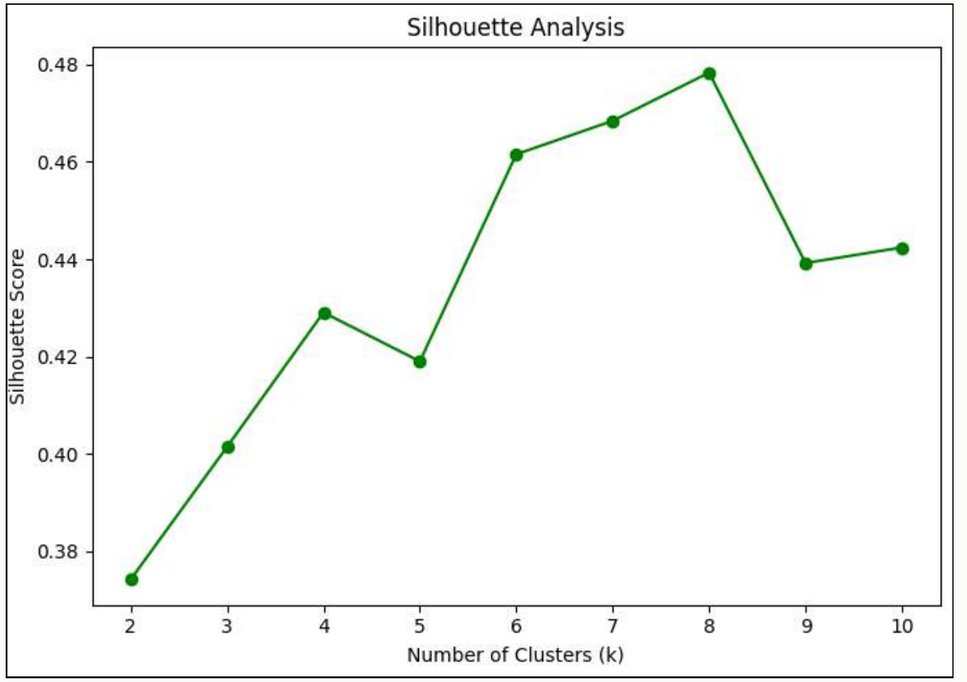
```
Standardization complete.

Original feature means:
 genre_code                4.272000
is_premium                 0.500000
hours_per_listener       464.239339
recent_stream_ratio        0.072394
dtype: float64

Scaled feature means (≈0):
 genre_code               -6.394885e-17
is_premium                0.000000e+00
hours_per_listener       -3.552714e-17
recent_stream_ratio       8.171241e-17
dtype: float64

Scaled feature std (≈1):
 genre_code                1.001002
is_premium                 1.001002
hours_per_listener         1.001002
recent_stream_ratio        1.001002
dtype: float64
```

K-MEANS Tune Model

- Elbow plot suggests k between 5–7



- Silhouette score favors higher k values



Overall silhouette values remain moderate, suggesting weak but usable cluster separation

- K value: 6, resulted in lower precision
- K Value Selected: 8

K-MEANS Tune Model

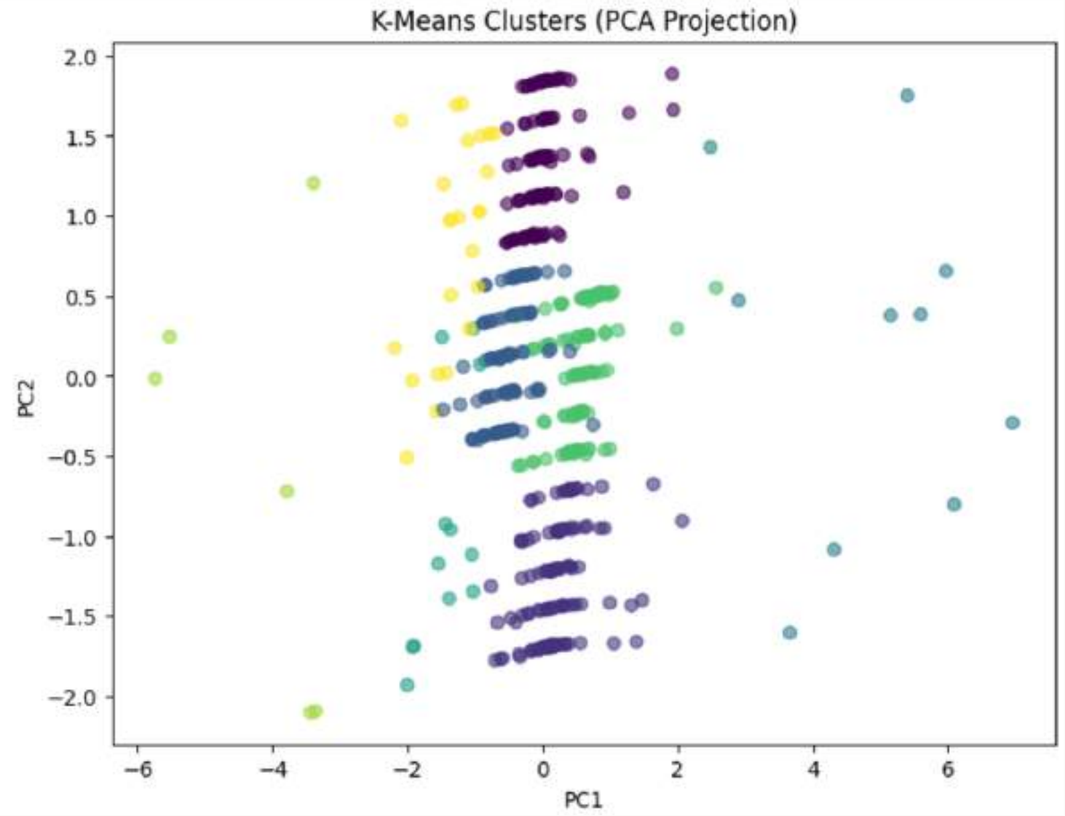
- Trained the K-Means model using 8 clusters
- Set max_iter = 300 to allow the algorithm to converge
- Used n_init = 10 to improve centroid initialization stability
- Fixed random_state = 42 to ensure reproducible results

K-MEANS Model Evaluation

```
K-Means Results
Clusters: 8
Silhouette Score: 0.478
Davies-Bouldin Index: 0.756
```

```
Cluster sizes:
0    103
1    123
2    116
3     10
4     13
5    105
6      6
7     24
Name: count, dtype: int64
```

K-MEANS



	genre_code	is_premium	hours_per_listener	recent_stream_ratio
PC1	0.235648	0.338162	0.706766	-0.574977
PC2	0.737429	-0.671619	0.019409	-0.068914

PC1 (Principal Component 1)

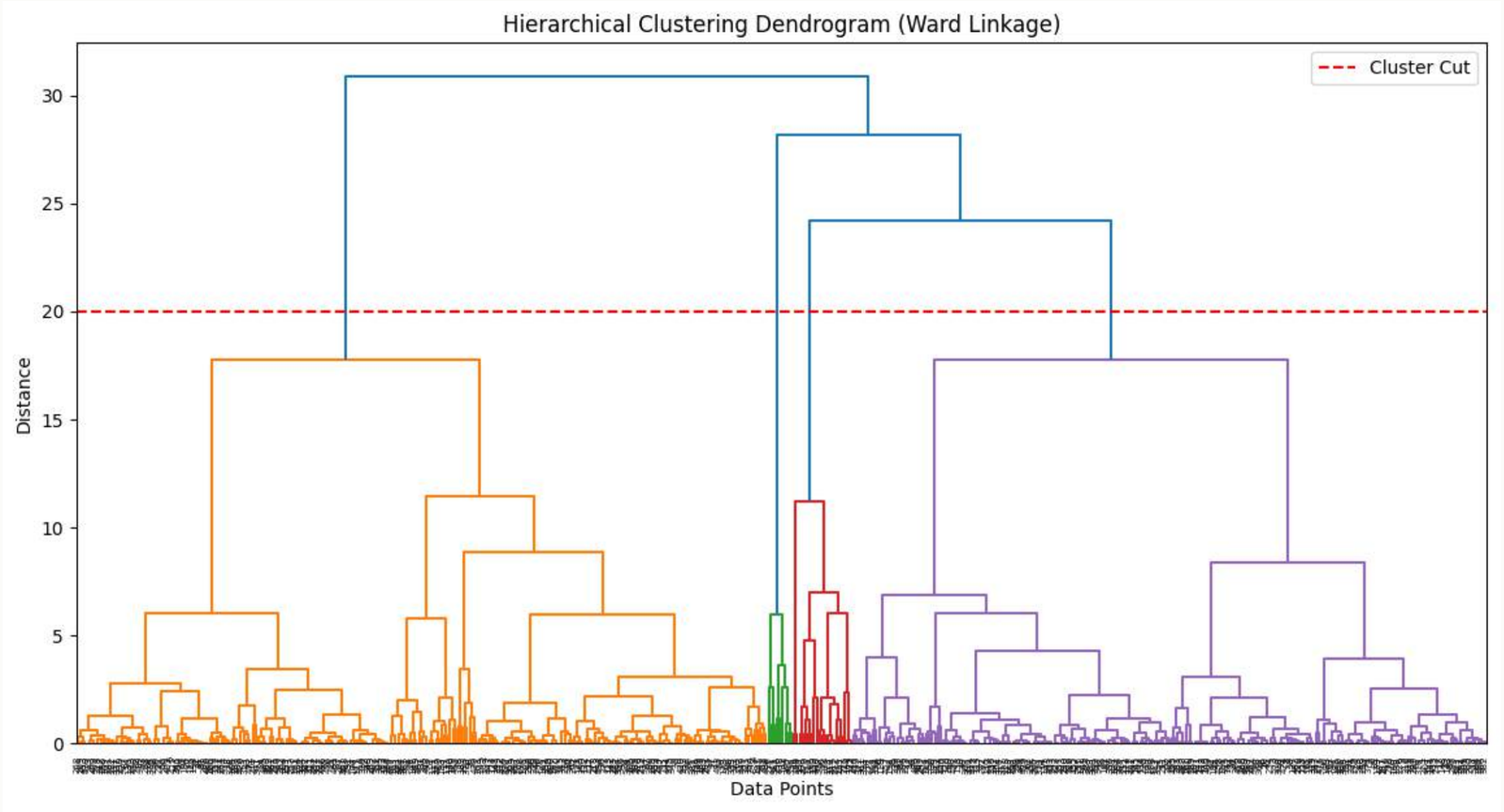
- PC1 mainly measures how much users listen overall
- Right side (high PC1) → heavy listeners, often premium, high total hours
- Left side (low PC1) → light users or recently inactive users
- Negative recent_stream_ratio means long-term heavy users may not be “recently active”

PC2 (Principal Component 2)

- PC2 separates users by content preference and subscription type
- Top (high PC2) → diverse genre listeners, mostly free users
- Bottom (low PC2) → premium users, more focused preferences

Hierarchical Clustering Dendogram

Tune Model



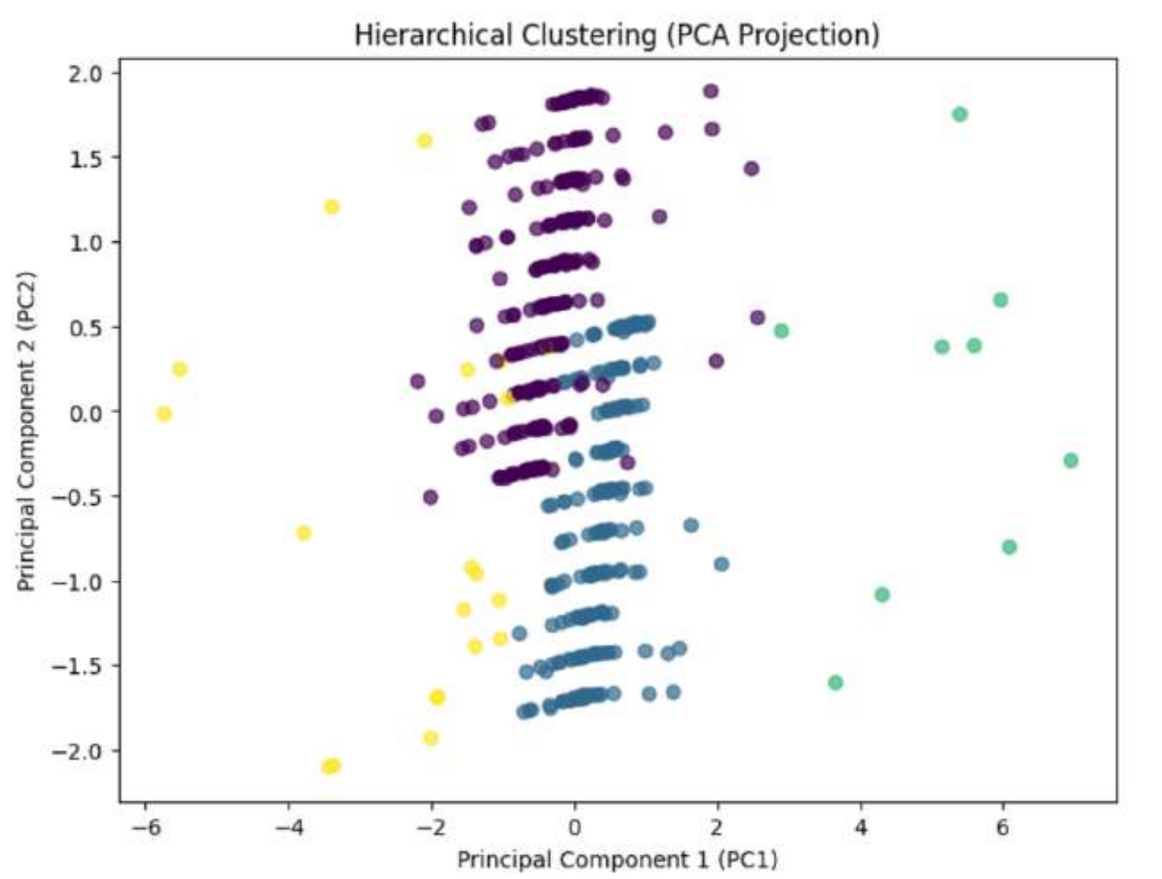
- This cut intersects four major branches, suggesting 4 natural clusters in the data.

Hierarchical Clustering Model Evaluation

- Applied Agglomerative Hierarchical Clustering with Ward linkage
- Selected 4 clusters based on dendrogram structure and cluster separation
- Used scaled features to ensure fair distance computation

```
Hierarchical Clustering Results
Silhouette Score: 0.424
Davies-Bouldin Index: 0.870
```

Applied Agglomerative Hierarchical Clustering



PC1 (Principal Component 1)

- Represents overall listening intensity
- Strongly influenced by hours per listener
- Separates heavy vs. light users

PC2 (Principal Component 2)

- Captures content and subscription behavior
- Influenced by genre preference and premium status
- Distinguishes user type and engagement style

Overlap suggests some users share similar listening patterns across clusters

	genre_code	is_premium	hours_per_listener	recent_stream_ratio
PC1	0.235648	0.338162	0.706766	-0.574977
PC2	0.737429	-0.671619	0.019409	-0.068914

Model Evaluation & Comparison

Evaluation Metrics Used

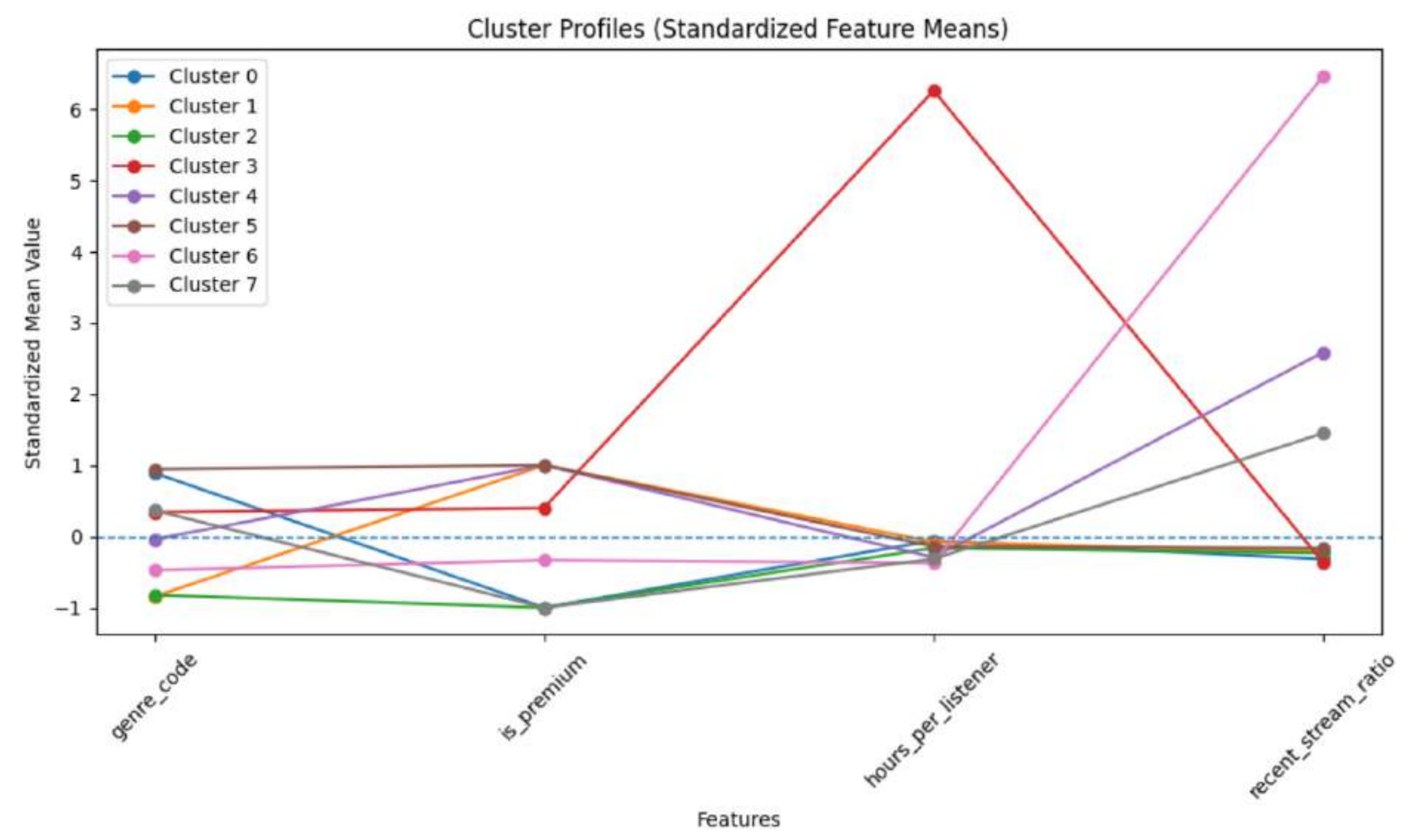
- Silhouette Score → measures cluster separation (higher is better)
- Davies–Bouldin Index → measures cluster compactness (lower is better)

Results

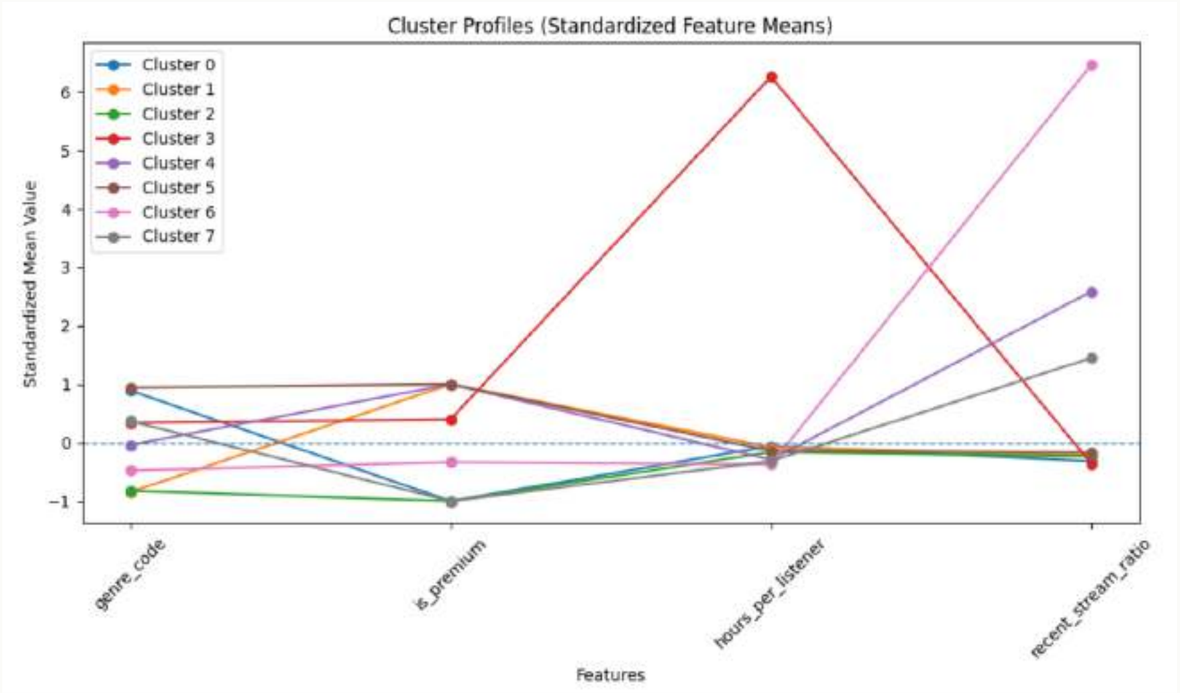
- K-Means
  - Silhouette Score: 0.478
  - Davies–Bouldin Index: 0.756
- Hierarchical Clustering
  - Silhouette Score: 0.424
  - Davies–Bouldin Index: 0.870

K-Means was selected as the final clustering model

Cluster Interpretation

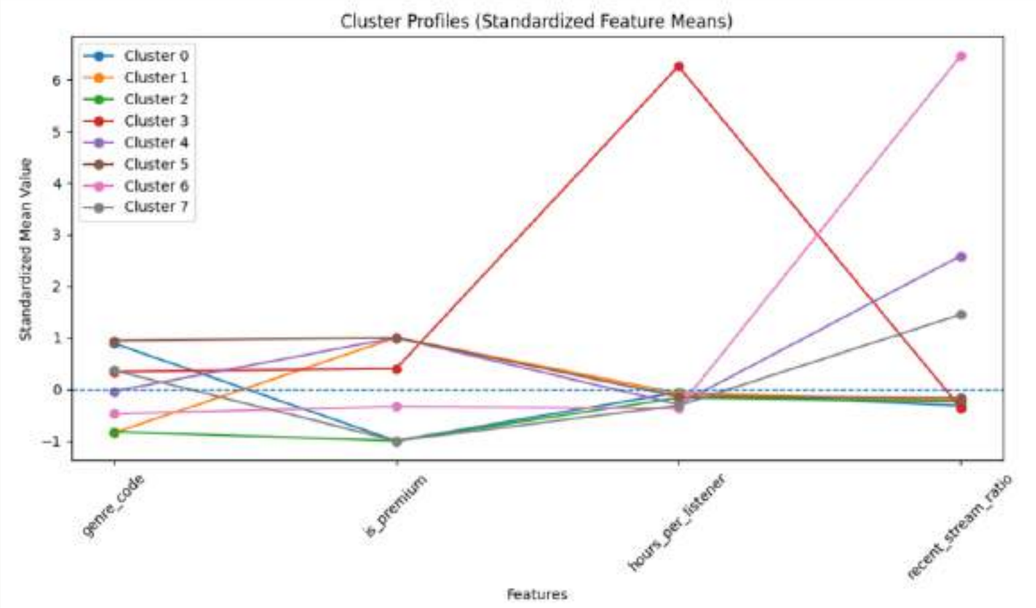


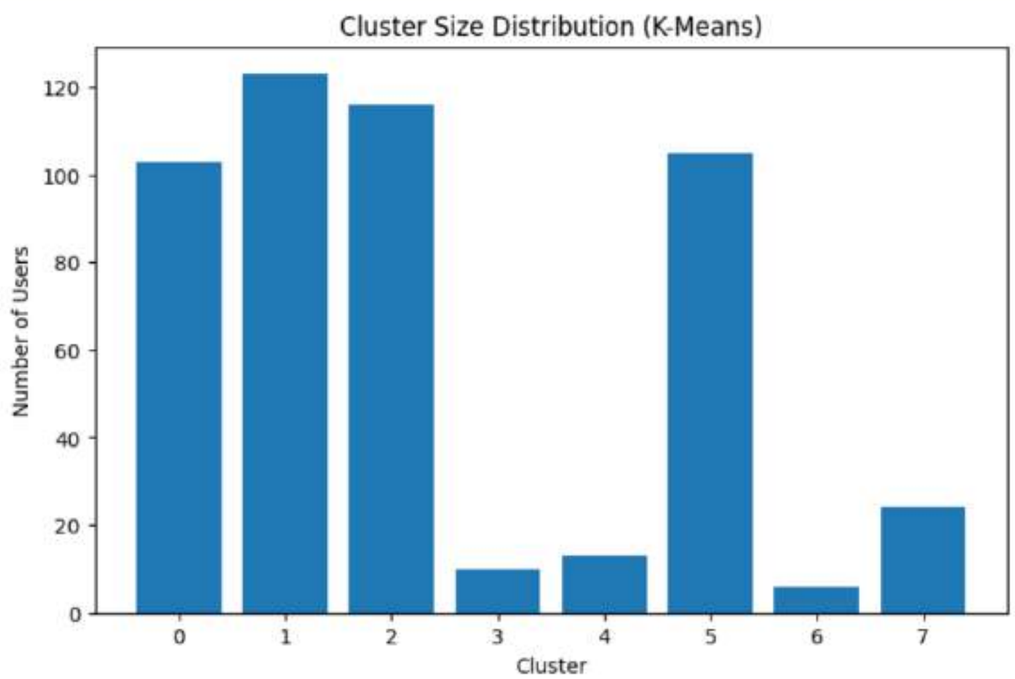
Selected Model: K-Means Cluster Interpretation



- Cluster 0: Free users with moderate listening hours and low recent activity
- Cluster 1: Premium users with balanced listening and steady recent engagement
- Cluster 2: Free users with moderate usage and average recent streaming
- Cluster 3: Heavy users with extremely high listening hours and stable engagement
- Cluster 4: Premium users with low listening hours but high recent activity
- Cluster 5: Premium users with diverse genres and moderate engagement
- Cluster 6: Very low-activity users with minimal listening and high churn risk
- Cluster 7: Free users with many genres and low hours but noticeable recent engagement spikes

Interpretation





- Most users belong to clusters 0, 1, 2, and 5
- Driven by regular users rather than extreme heavy or inactive users
- A large free-user base presents strong free-to-premium conversion opportunities
- Premium users already show steady engagement, supporting reliable subscription revenue
- User behavior is predictable, making retention and personalization more impactful than targeting edge cases

Business Goal Check

The clustering model successfully groups users based on listening behaviour and activity patterns

Purpose

- Target free users (Clusters 0 & 2) with personalized upgrade offers based on their listening habits
- Retain premium users (Clusters 1 & 5) through loyalty perks, exclusive content, and recommendations