Moosic playlists Strategy

Aidar, Elaheh, Aida

Key Metrics for Evaluation

Data Overview

Datasets:
- the 5235 song dataset from Spotify

Data preparation:
- Duplicates
- Feature selection
- Outliers

Final set:
- 5164 songs dataset

Songs features:

☑ name
☑ artist
☑ danceability
☑ energy
  key
☑ loudness
☑ mode
☑ speechiness
☑ acousticness
☑ instrumentalness
☑ liveness
☑ valence
☑ tempo
  type
☑ duration_ms
  time_signature
  id
  html

Key Questions

🎶 **Are Spotify's audio features able to identify "similar songs", as defined by humanly detectable criteria?**
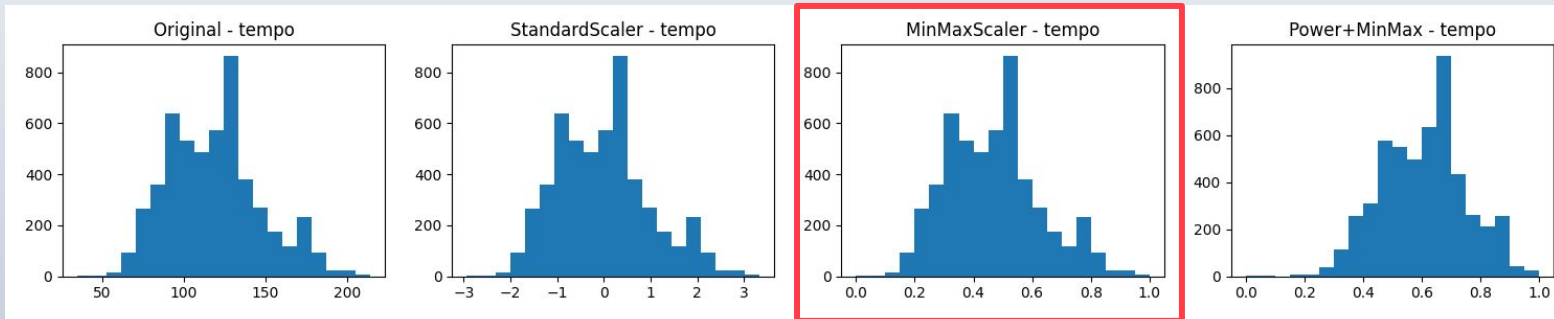
🎶 **Is K-Means a good method to create playlists?**

Feature scaling
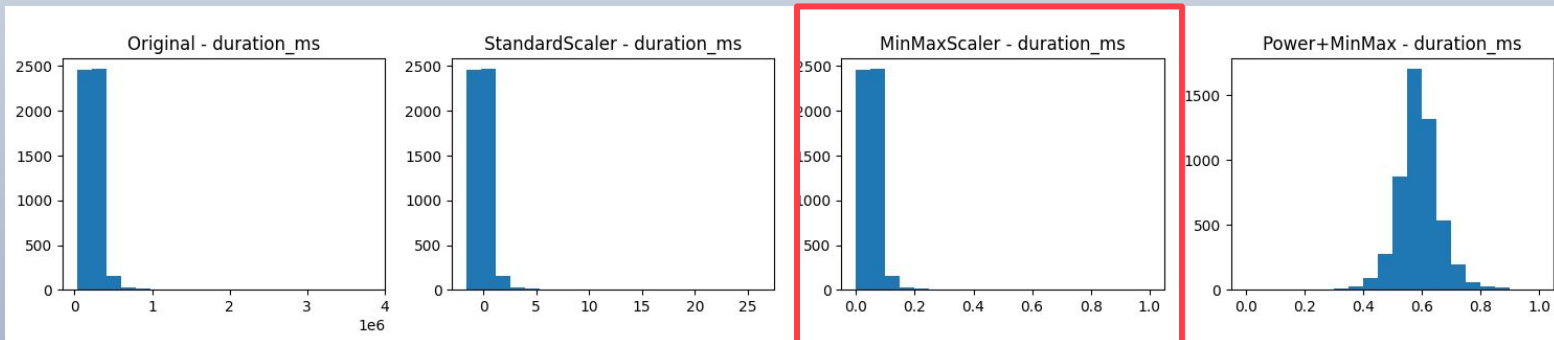
🎵 Features like **tempo** and **duration_ms** have different ranges.

🎵 Scaling ensures equal contribution to clustering → **MinMaxScalar**

Dimensionality Reduction Using PCA

PCA (Principal component analysis):

- ▷ A **dimensionality reduction** technique.
- ▷ Captures the most important patterns or variance in the data.

Cumulative explained variance:

- ▷ Retained **95% of the variance** to ensure high data quality for clustering.
- ▷ Reduced 11 audio features to **7 principal components**.

Cumulative Explained Variance
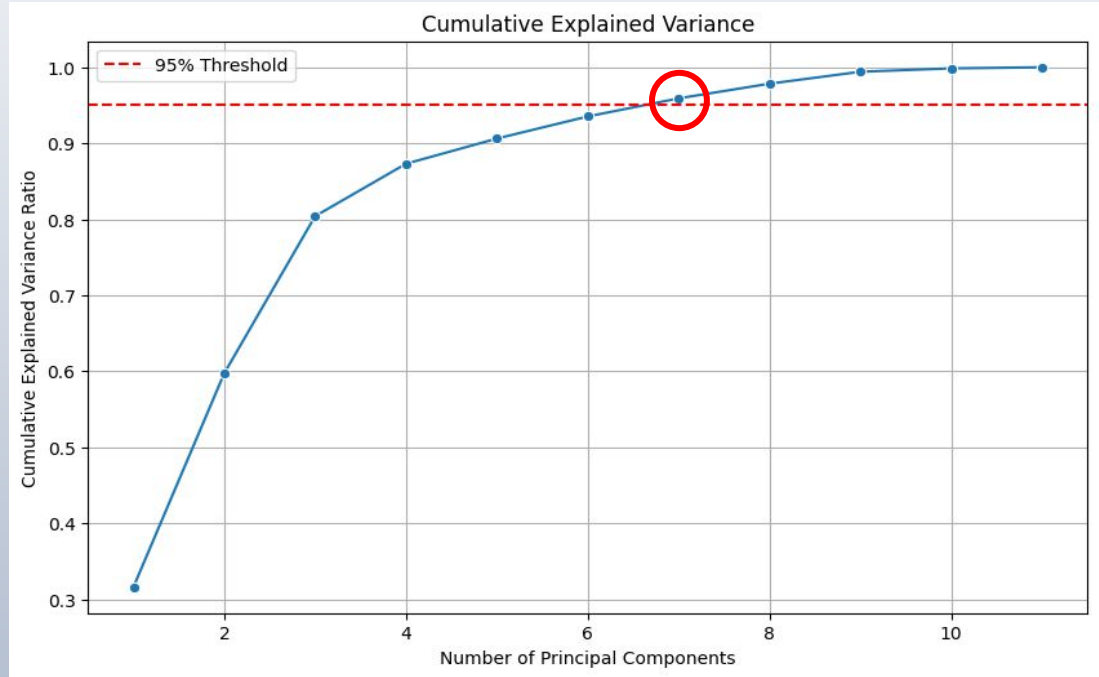
K-Means Clustering – Optimal Number of Clusters

▷ Automatically organizes data into clusters based on similarity.

▷ **Technical Criterion:** Evaluated using the **Silhouette Score**, which measures how well data points fit within their clusters.

▷ **Business Constraint:** Playlists typically range from **50 to 250 songs**, ensuring customer satisfaction and usability.
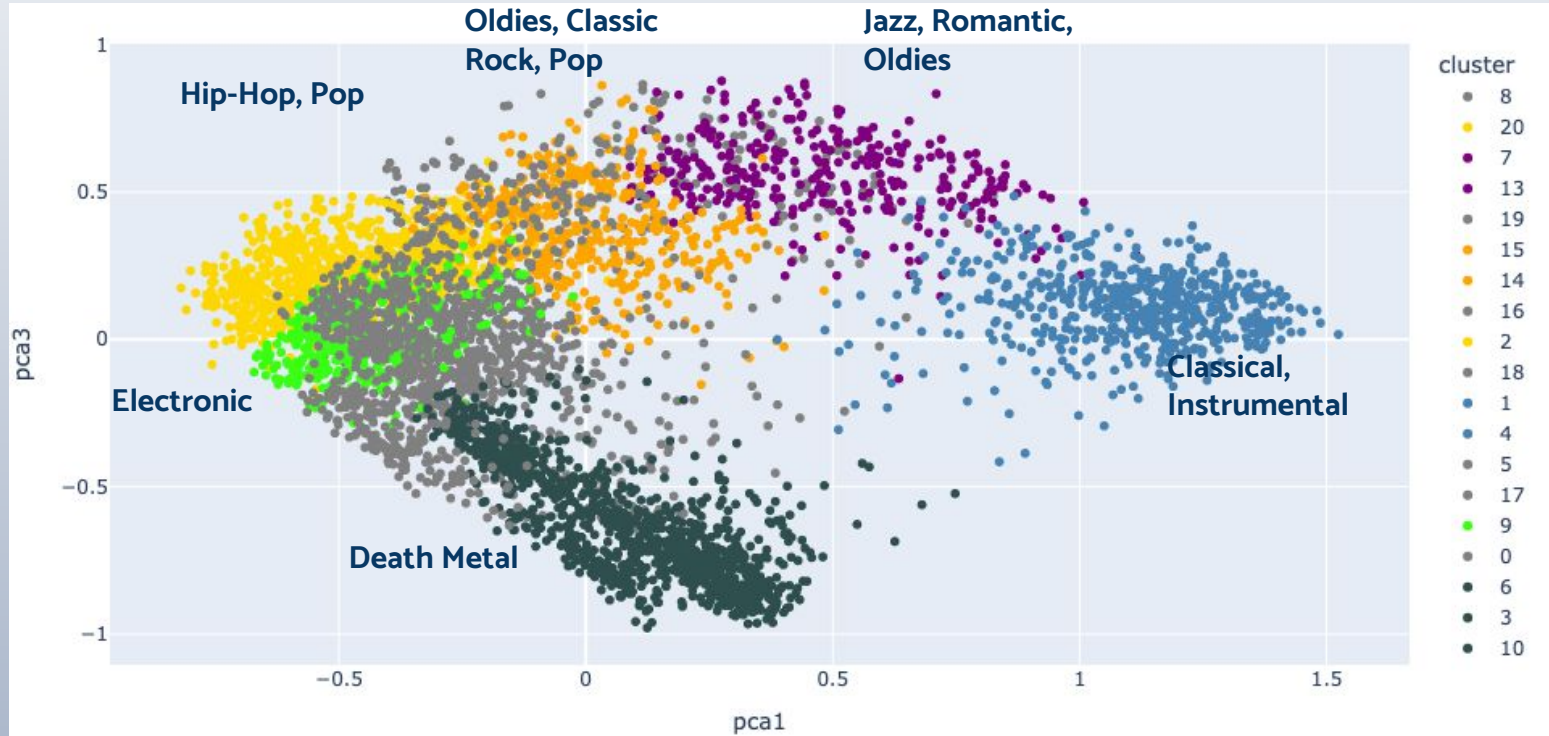

Silhouette Score from 2 to 49 Clusters
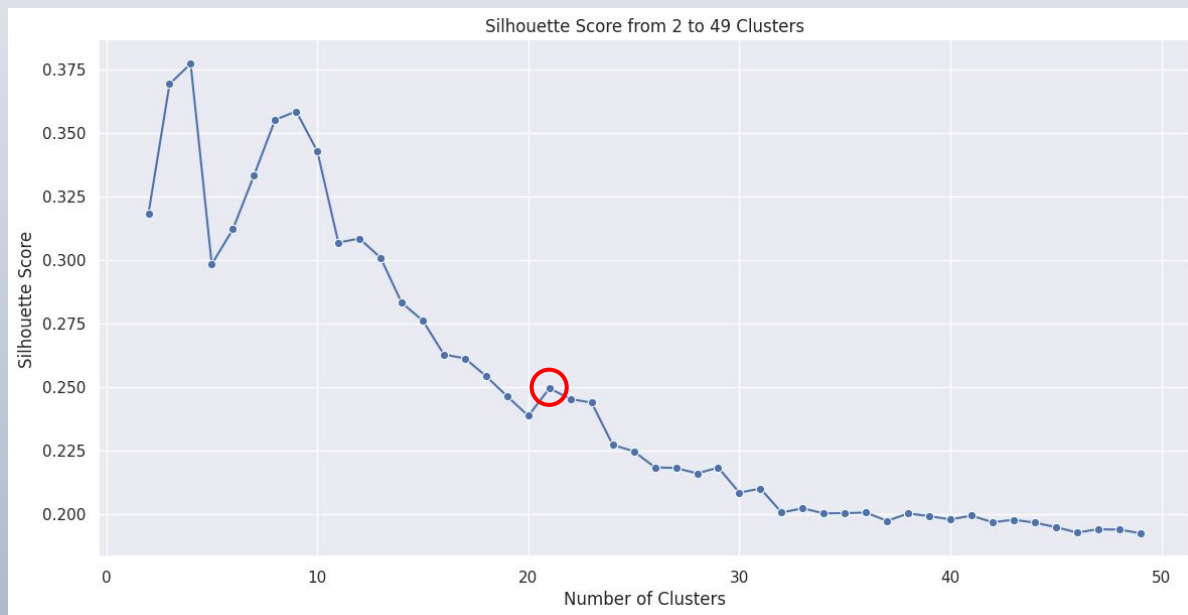
Final Choice:

Based on both business and technical factors, we selected **21 clusters** (adjust based on actual result).
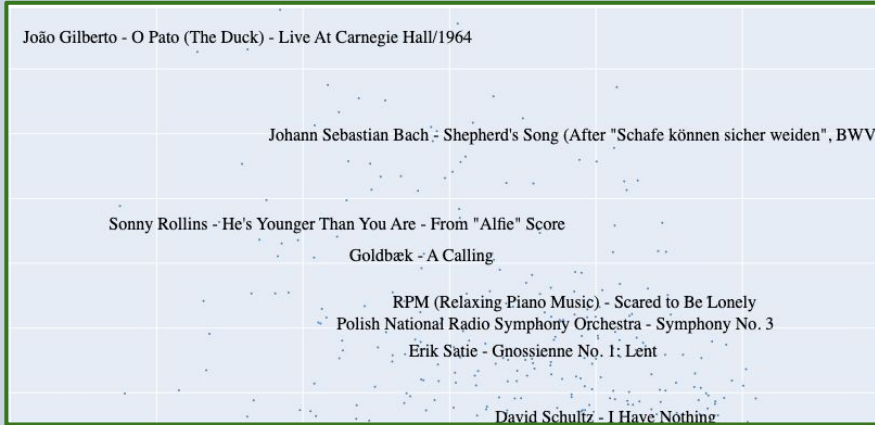
K-Means Clustering – Results

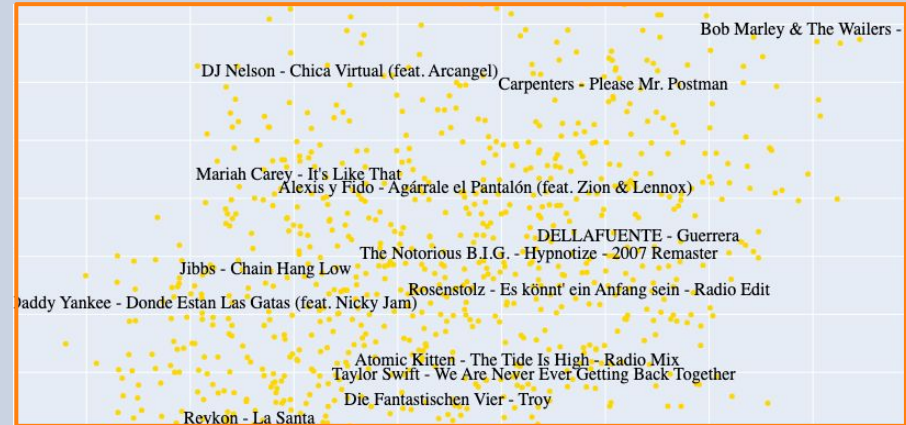Projection of our songs onto a **2-dimensional** component space

K-Means Clustering – Example Playlist

Random samples from **Classical** / **Instrumental** Playlist

Random samples from **Hip-Hop / Pop** Playlist

João Gilberto - O Pato (The Duck) - Live At Carnegie Hall/1964

Johann Sebastian Bach - Shepherd's Song (After "Schafe können sicher weiden", BWV

Sonny Rollins - He's Younger Than You Are - From "Alfie" Score

Goldbæk - A Calling

RPM (Relaxing Piano Music) - Scared to Be Lonely

Polish National Radio Symphony Orchestra - Symphony No. 3

Erik Satie - Gnossienne No. 1: Lent

David Schultz - I Have Nothing

Bob Marley & The Wailers -

DJ Nelson - Chica Virtual (feat. Arcangel)

Carpenters - Please Mr. Postman

Mariah Carey - It's Like That

Alexis y Fido - Agárrale el Pantalón (feat. Zion & Lennox)

DELLAFUENTE - Guerrera

The Notorious B.I.G. - Hypnotize - 2007 Remaster

Jibbs - Chain Hang Low

Rosenstolz - Es könnt' ein Anfang sein - Radio Edit

Daddy Yankee - Donde Estan Las Gatas (feat. Nicky Jam)

Atomic Kitten - The Tide Is High - Radio Mix

Taylor Swift - We Are Never Ever Getting Back Together

Die Fantastischen Vier - Troy

Reykon - La Santa

Conclusion

Results:

- Used scaling, PCA, and K-Means to create **21 playlists** with clear themes (e.g., hip-hop, death metal, instrumental).
- Spotify's audio features can generally identify similar songs, with only minor errors.
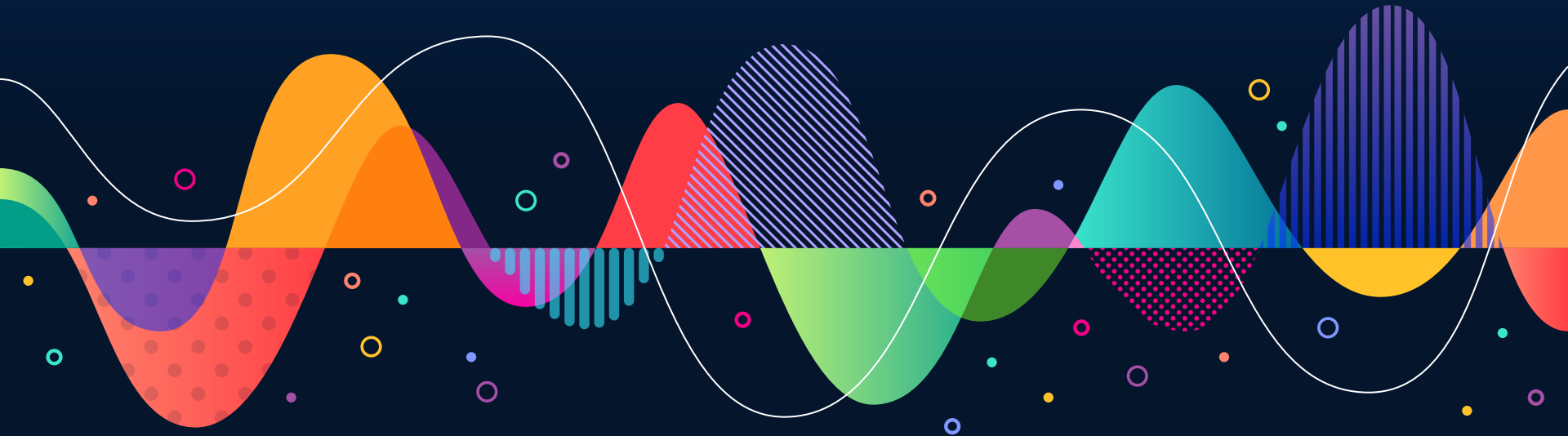- K-Means works well, but it's the only method we tested so far!

Next steps:

- Spotify's features are strong but need additional data to better capture human perceptions.
- Explore new algorithms and advanced techniques for improved personalization.
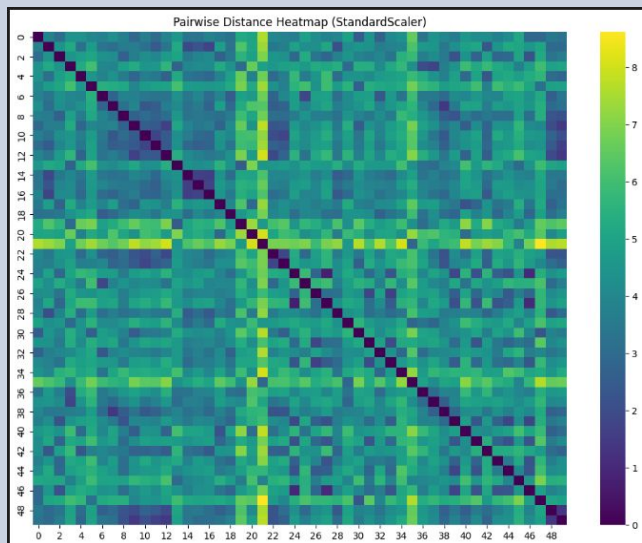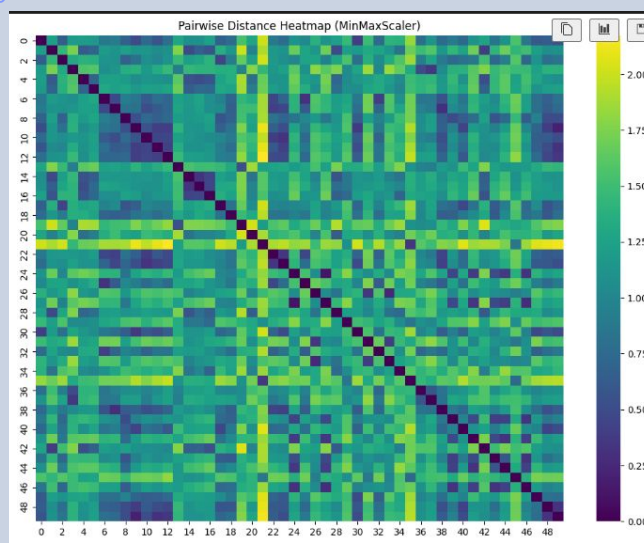
THANKS!

Backup slides

Feature scaling (extra slide)

Pairwise Distance Heatmap (**StandardScaler**)

Pairwise Distance Heatmap (**MinMaxScaler**)

Feature scaling (extra slide)