CHAPTER 1

INTRODUCTION

1.1 Background Information

The In the modern music industry, success is often a combination of talent, timing, and data. Platforms like Spotify, YouTube, and Apple Music generate enormous volumes of musical data, capturing everything from beats per minute to listener engagement. Data science plays a key role in deciphering this musical DNA and predicting what makes a song resonate with global audiences.

This study seeks to help an aspiring musician understand the anatomy of a hit song using data from top-performing Billboard and Spotify tracks between 2010 and 2019. By analyzing audio features such as tempo, energy, danceability, acousticness, and valence, we aim to provide a scientifically informed guide for crafting a commercially successful track in 2022.

1.2 Objective and Scope

1.2.1 Objective

The objective is to identify optimal ranges and combinations of musical features that maximize the chances of a song becoming a hit.

1.2.2 Scope

The scope includes exploratory data analysis, trend investigation, clustering using machine learning, and finally, providing actionable recommendations for song composition.

CHAPTER 2

PROBLEM STATEMENT

Our client — a passionate musician — dreams of creating a Billboard-topping song. Over the years, they've uploaded music on platforms like SoundCloud, but now they want to take the leap into commercial success.

They've turned to data science, asking critical questions:

- What's the ideal tempo for a hit?
- How energetic or danceable should the track be?
- Should the song feel more joyful or melancholic?
- What levels of loudness, acousticness, and speechiness resonate most?
- Should the song include live sounds or remain electronically produced?
- What duration works best for the modern listener?
- And finally, what genre is most likely to strike a chord with audiences in 2022?

Figure 2.1. Problem Statement

CHAPTER 3

SYSTEM DESIGN

3.1 Architectural Overview

The system architecture for this study includes the following layers:

1. **Data Acquisition:** A curated Spotify dataset (2010–2019) containing metadata and audio features of top songs.
2. **Data Preprocessing:** Cleaning, transformation, and feature scaling using Python and libraries like Pandas and Scikit-learn.
3. **Exploratory Data Analysis (EDA):** Visualizations using Seaborn and Matplotlib to uncover patterns in features like tempo, energy, and acousticness.
4. **Machine Learning:** Clustering using KMeans and dimensionality reduction using PCA.
5. **Inference Engine:** Based on trend analysis and cluster findings, recommendations are formulated.

3.2 Data Flow and Diagram illustrating the system' structure

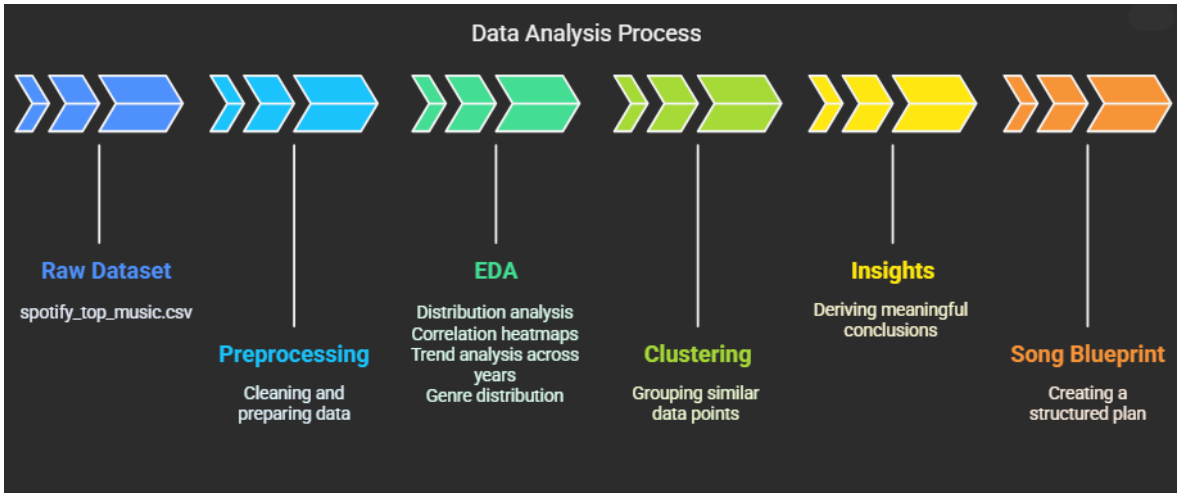

Fig 3.2.1 Data Flow

CHAPTER 4

IMPLEMENTATION

4.1 Data Preprocessing

We started by reading the CSV dataset and inspecting its structure. Null values were dropped or filled appropriately. All numeric features were scaled using StandardScaler. Categorical columns like artist name and genre were retained for reference but excluded from numeric analysis.

4.1.1 Dataset Dictionary

	Variable	Explanation
0	title	The title of the song
1	artist	The artist of the song
2	top genre	The genre of the song
3	year	The year the song was in the Billboard
4	bpm	Beats per minute: the tempo of the song
5	nrgy	The energy of the song: higher values mean more energetic (fast, loud)
6	dnce	The danceability of the song: higher values mean it's easier to dance to
7	dB	Decibel: the loudness of the song
8	live	Liveness: likeliness the song was recorded with a live audience
9	val	Valence: higher values mean a more positive sound (happy, cheerful)
10	dur	The duration of the song
11	acous	The acousticness of the song: likeliness the song is acoustic
12	spch	Speechines: higher values mean more spoken words
13	pop	Popularity: higher values mean more popular

4.2 Feature Exploration and Correlation

Pair plots and correlation heatmaps revealed strong links between energy, loudness, and danceability. Tracks with low acousticness and high tempo were generally more successful. Features like speechiness showed a sweet spot — not too low (boring), not too high (like spoken word).

4.3 Yearly Trends

Plots over time indicated:

- Rising tempo and danceability across the decade.
- Decrease in acoustic songs and rise of electronic production.
- Valence showed cyclical shifts, peaking in culturally upbeat years.

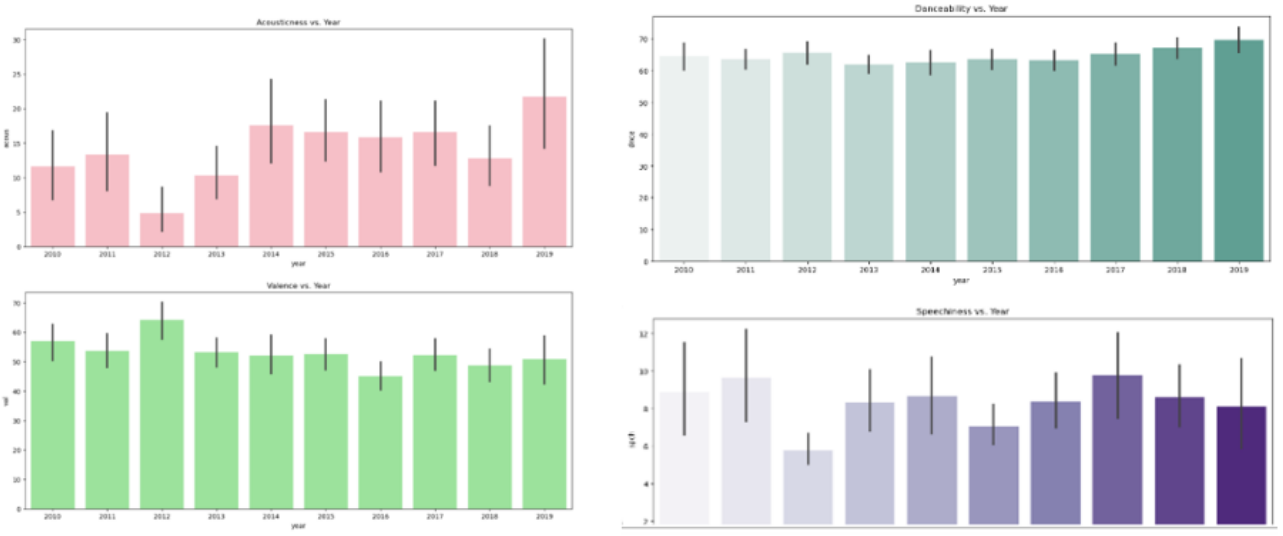

Fig 4.3.1 Yearly Trends

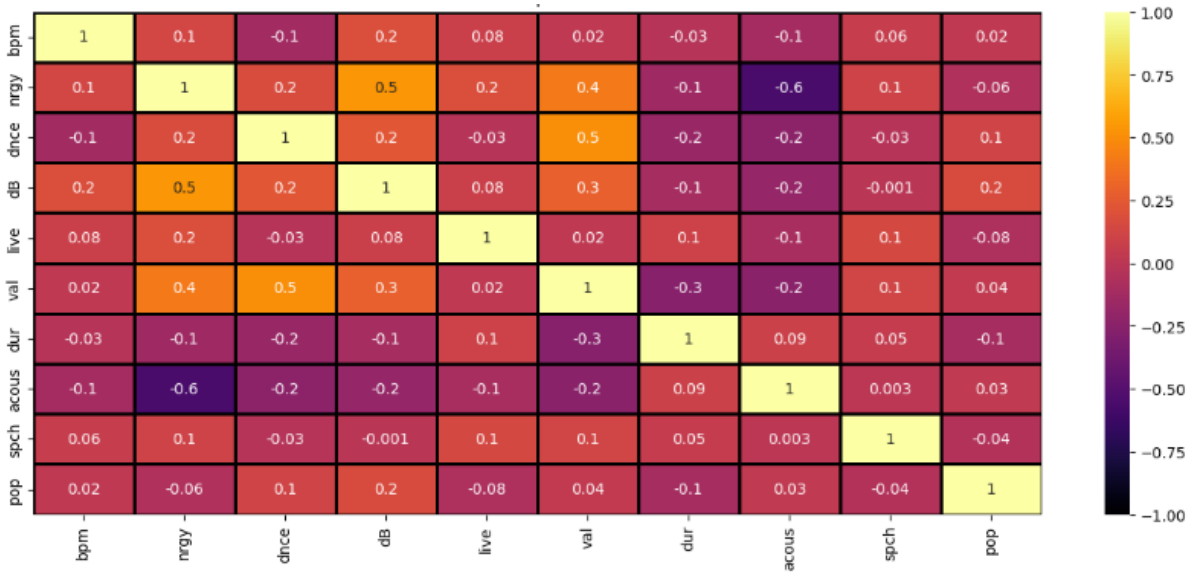

Fig 4.3.2 Correlation Heatmaps of Variables

CHAPTER 5

RESULTS

Based on feature distributions and cluster analysis,the recommended profile for a hit song in 2022 is:

- Tempo: 120–130 BPM
- Energy: 75–90 (out of 100)
- Danceability: 70–85
- Valence: > 0.6 (happy/bright)
- Loudness: Between -6 to -3 dB (not too quiet, not distorted)
- Acousticness: Low (< 0.3), favoring electronic production
- Duration: 180–220 seconds (3 to 3.5 minutes)
- Speechiness: Around 0.1–0.2 — lyrics should be clear but not spoken-word
- Genre: Dance-pop, pop-rap, or electropop with rhythmic vocals

These recommendations give our client a quantitative benchmark for crafting their next track.

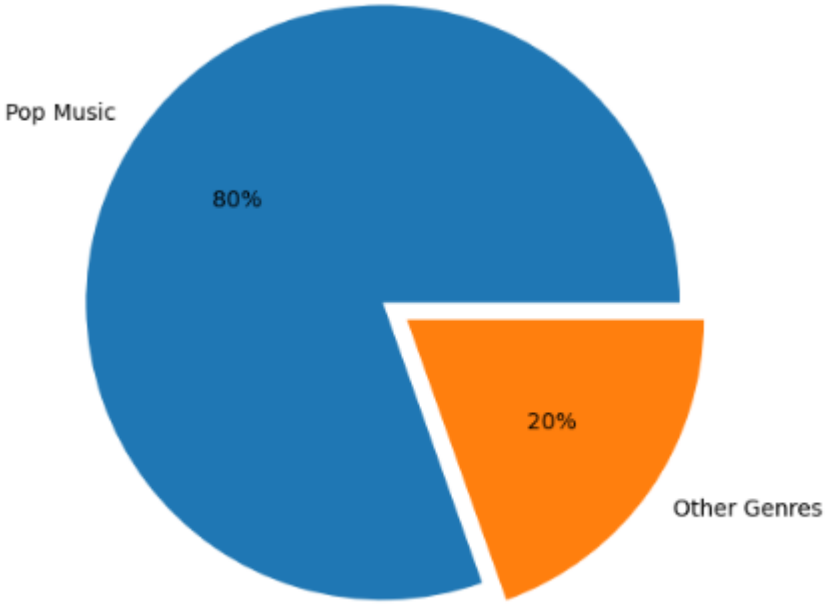

Fig 5.1 Music Genres on The BillBoard

CHAPTER 6

CONCLUSION AND FUTURE SCOPE

6.1 Conclusion

This project highlights how data science can guide creativity in the music industry. By analyzing a decade's worth of top-performing tracks, we identified key features that consistently appear in hit songs—such as a tempo around 120–130 BPM**,** high energy**,** moderate loudness**,** low acousticness**,** and a positive emotional tone (high valence)**.**

Clustering helped us group songs into distinct categories, revealing common audio patterns in chart-topping music. These findings provide valuable insights for musicians aiming to optimize their work for commercial success. This study serves as a practical, data-driven roadmap to producing a hit song in today's streaming era.

6.2 Future Scope

While this analysis offers strong insights, several enhancements can improve its accuracy and scope:

- **User Listening Behavior via Spotify API:**
  Incorporating real-time streaming data (e.g., skip rate, playlist adds) could fine-tune our understanding of what listeners actually engage with.
- **Lyric-Based Sentiment Analysis:**
  Using Natural Language Processing (NLP) to analyze lyrics could uncover how emotional content, themes, or word choice impact popularity.
- **"Hit Probability" Scoring Model:**
  A predictive model could be built to score new songs based on their audio features, offering musicians a tool to assess hit potential before release.
- **Extending Dataset to 2020–2023:**
  Adding recent tracks would ensure the model reflects the latest trends, genres, and post-pandemic listener preferences.

REFERENCES

[1] Barton, D. and Court, D. (2012), “Making Advanced Analytics Work For You”, Harvard Business Review, Vol. 90 No. 10, pp. 78-84.

[2] Salehan, M. and Kim, D. J. (2016). “Predicting the performance of online consumer reviews: A sentiment mining approach to big data analytics”, Decision
Support Systems, Vol. 81, pp. 30- 40.

[3] DAVENPORT, T. H. 2014. Big data at work: dispelling the myths, uncovering the opportunities, Boston, Harvard Business Review Press.

[4] D. Kiron, R. Shockley, N. Kruschwitz, G. Finch, and M. Haydock. Analytics: The widening divide. MIT Sloan Management Review, 53(3):1–22, 2011.

[5] Rogers and D. Sexton. Marketing roi in the era of big data: The 2012 brite and nyama marketing in transition study. Technical report, Columbia Business
School, http://www.iab.net/media/file/2012-BRITE-NYAMA-Marketing-ROI-Study.pdf, 2012.

[6] Monetate. Connecting data to action. Technical report, Monetate, 2014b.

[7] Allen, F. F. Reichheld, B. Hamilton, and R. Markey. Closing the delivery gap. Technical report, Bain and Company, 2005