Decoding The Hits :

Analytics Behind Spotify's Popular Tracks

Submitted by :
Shreya Mishra

Play. Pause. Predict.

OBJECTIVE

Predict whether a song will be popular (popularity ≥ 65) to improve Spotify's recommendation system.

SCOPE

- Clean and preprocess Spotify song data
- Build and compare predictive models
- Optimize predictions based on profit
- Analyze patterns in song traits and genres
- Provide actionable insights for better playlist curation

Data Pre-Processing

artist	499
song	499
song_name_len	499
duration_ms	499
explicit	499
year	499
popularity	499
hot	499
danceability	499
energy	499
key	499
loudness	499
mode	499
speechiness	499
acousticness	499
instrumentalness	499
liveness	499
valence	499

artist	0
song	0
song_name_len	0
duration_ms	0
explicit	0
year	0
popularity	0
hot	0
danceability	0
energy	0
key	0
loudness	0
mode	0
speechiness	0
acousticness	0
instrumentalness	0
liveness	0
valence	0
tempo	0

Handling Missing Values

Handled missing data by filling all null values with zeros for consistent analysis.

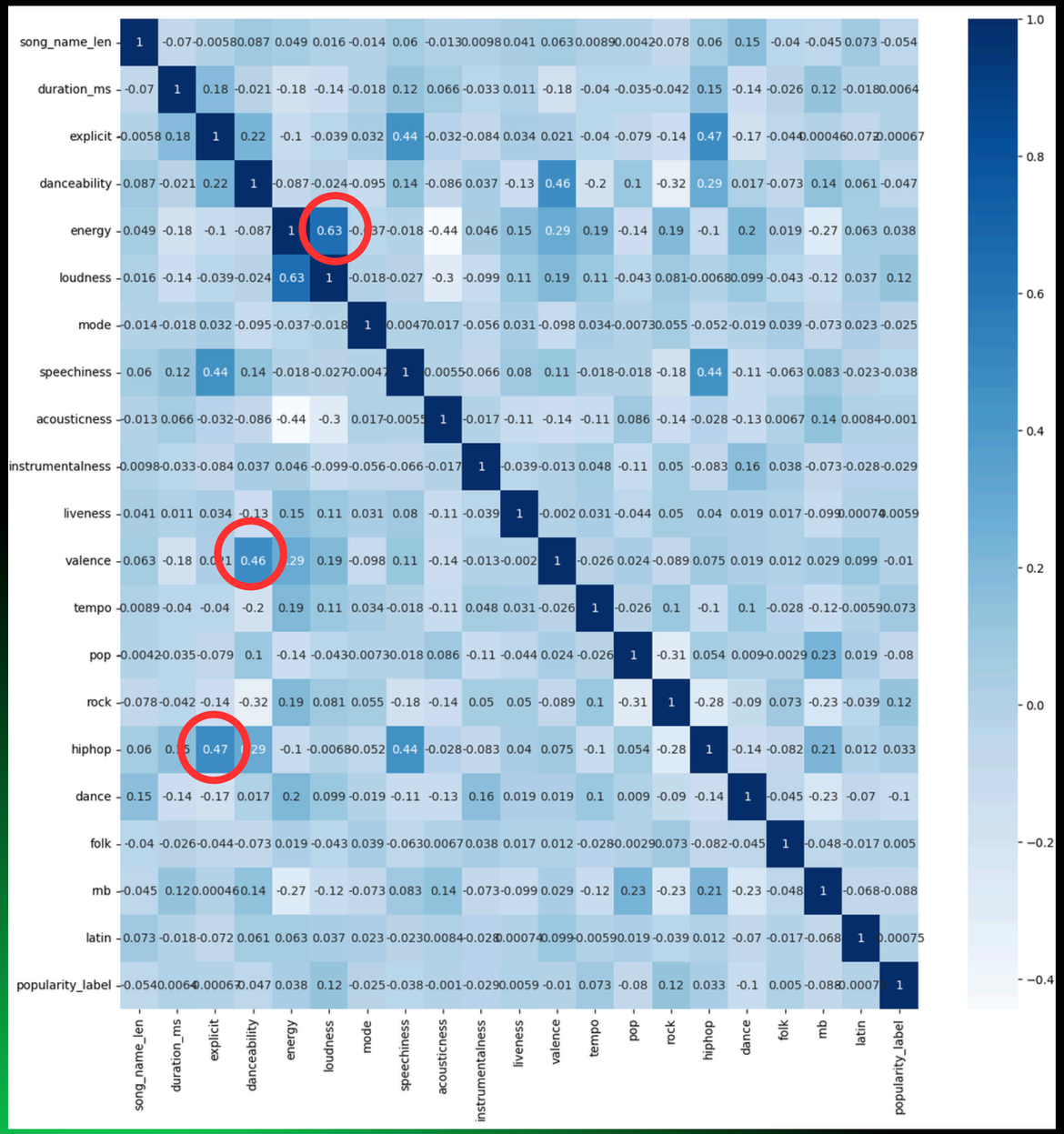
Selecting relevant attributes

Selected relevant features for modeling by retaining impactful numerical, categorical, and genre-based variables. Applied encoding where needed and prepared the dataset for predictive analysis.

Creating outcome variable

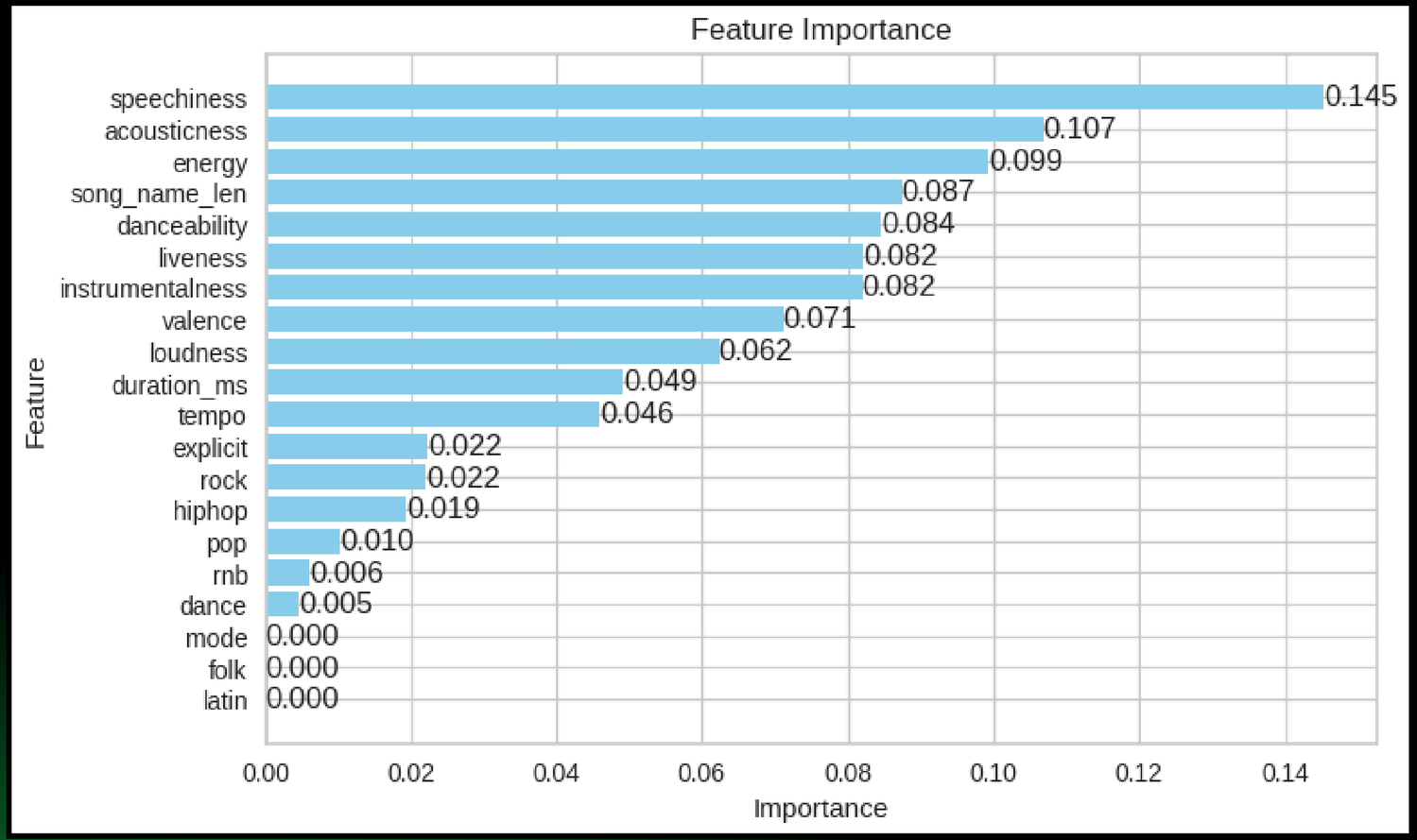
Created the outcome variable by categorizing songs based on their popularity score to enable binary classification.

Correlation Matrix



The correlation matrix highlights strong relationships like energy–loudness (0.63), valence–danceability (0.46), and hiphop–danceability (0.47), revealing key traits of popular songs.

Feature Importance



Feature importance analysis reveals that speechiness (0.145), acousticness (0.107), and energy (0.099) are the strongest predictors of song popularity, suggesting that vocal presence, acoustic elements, and intensity significantly influence user preferences.

Question 1

The company has commissioned you to develop a predictive model to enhance its recommender systems by focusing on popular songs. The primary goal of this project is to predict whether a song's popularity score exceeds 64 (>=65), enabling Spotify to prioritize songs that are more likely to engage users effectively.

Methodology followed:

- Developed three models:
  - Decision Tree
  - K-Nearest Neighbor
  - Logistic Regression

- Fine-tuned the models using 10-fold cross validation
- Evaluated the metrics such as accuracy

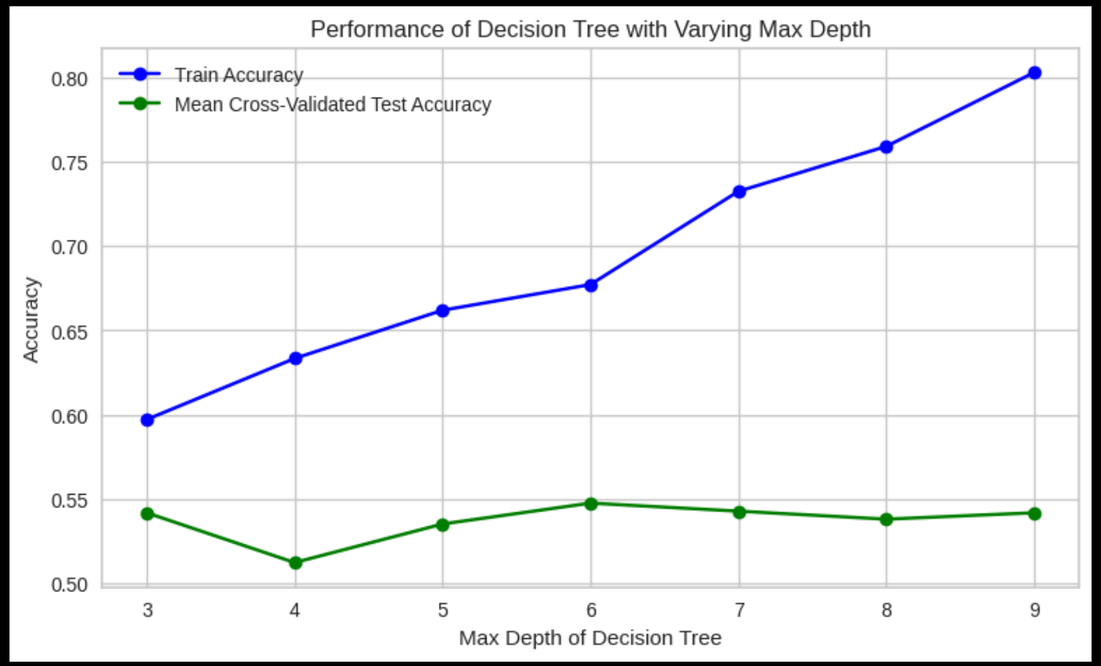
Decision Tree

Tested varying max_depth values (3–9) to observe model performance across complexity levels.

Best parameters found: max_depth=6, min_samples_split=7, splitter=best, with a mean CV test score of 0.5438.

Train accuracy dropped from 0.8895 to 0.7295 after tuning, showing reduced overfitting.

Test accuracy slightly decreased from 0.5800 to 0.5400, indicating a trade-off post-optimization.



	max_depth	min_samples_split	splitter	mean_CV_test_score
6	7	5	best	0.5438
0	5	5	best	0.5381

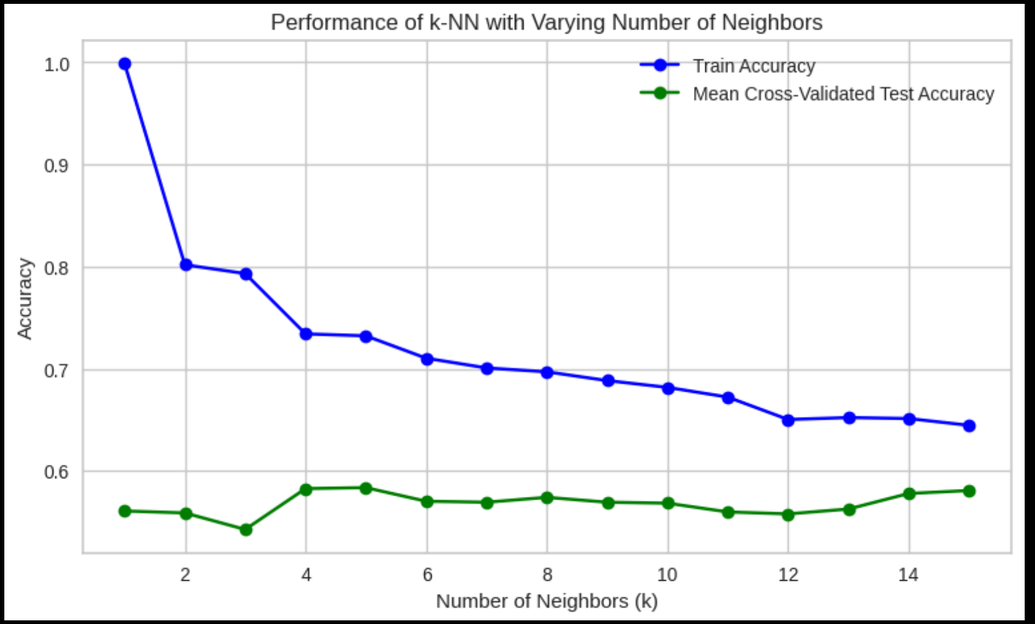
	Evaluation Metric	Original Decision Tree	Tuned Decision Tree
0	Train Accuracy	0.8895	0.7295
1	Test Accuracy	0.5800	0.5400
2	Precision	0.5899	0.5561
3	Recall	0.5614	0.4561
4	F1 Score	0.5753	0.5012

K-Nearest Neighbor

```
KNeighborsClassifier
KNeighborsClassifier(metric='euclidean', n_neighbors=3)
```

Before Fine-Tuning ·····→

	Evaluation Metric	Value
0	Train Accuracy	0.7933
1	Test Accuracy	0.5289
2	Recall	0.4649
3	Precision	0.5408
4	F1 Score	0.5000



	metric	n_neighbors	weights	mean_CV_test_score
3	euclidean	5	distance	0.5838
2	euclidean	5	uniform	0.5838

Fine-tuning improved test accuracy and overall model stability.

Train accuracy dropped, indicating reduced overfitting.

After Fine-Tuning ·····→

	Evaluation Metric	Original k-NN	Tuned k-NN
0	Train Accuracy	0.7933	0.7324
1	Test Accuracy	0.5289	0.5444
2	Precision	0.5408	0.5578
3	Recall	0.4649	0.4868
4	F1 Score	0.5000	0.5199

Logistic Regression

	feature	VIF
0	song_name_len	3.4762
1	duration_ms	33.4754
2	explicit	2.0653
3	danceability	30.9819
4	energy	33.5476
5	loudness	11.8478
6	mode	2.3184
7	speechiness	3.0135
8	acousticness	1.8320
9	instrumentalness	1.1178
10	liveness	2.7697
11	valence	12.0742
12	tempo	20.9781
13	pop	6.3900
14	rock	1.5776
15	hiphop	2.6260
16	dance	1.4961
17	folk	1.0408
18	rnb	1.7641
19	latin	1.0794
20	popularity_label	2.0125

Multicollinearity check →

	feature	VIF
0	song_name_len	3.2165
1	explicit	1.9584
2	mode	2.1681
3	speechiness	2.9395
4	acousticness	1.5411
5	instrumentalness	1.0732
6	liveness	2.5322
7	pop	4.9529
8	hiphop	2.5490
9	rock	1.3350
10	folk	1.0313
11	dance	1.4152
12	rnb	1.6865
13	latin	1.0774
14	valence	5.7734
15	popularity_label	1.8743

WITHOUT CROSS VALIDATION

	Evaluation Metric	Value
0	Train Accuracy	0.6362
1	Test Accuracy	0.5956
2	Recall	0.4825
3	Precision	0.6322
4	F1 Score	0.5473

→

Performance Metrics With Cross-Validation:

	Evaluation Metric	Value
0	Accuracy	0.5780
1	Recall	0.4986
2	Precision	0.5642
3	F1 Score	0.5294

Confusion Matrix - Cross Validation:
[[511 275]
[358 356]]

WITH CROSS VALIDATION

Managerial Insights & Recommendation

- **Focus on speech-driven, acoustic, and energetic songs**

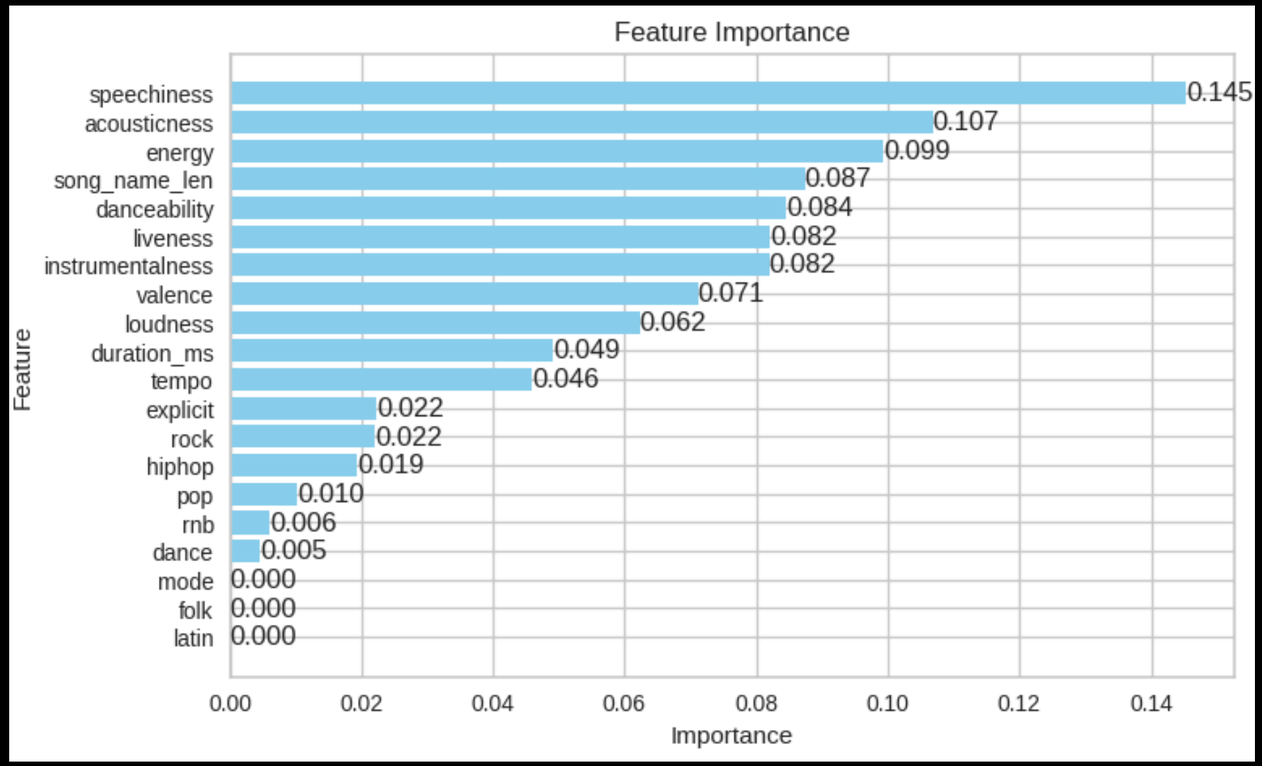
These are the top features driving popularity—songs with expressive vocals, acoustic quality, and high intensity are more likely to succeed.

- **Prioritize tracks with shorter names and higher danceability**

Danceability and concise song names positively impact engagement and should guide playlist curation.

- **Prioritize tracks with shorter names and higher danceability**

Cheerful, positive, and lively songs perform better—ideal for feel-good and party themed playlists.



Question 2

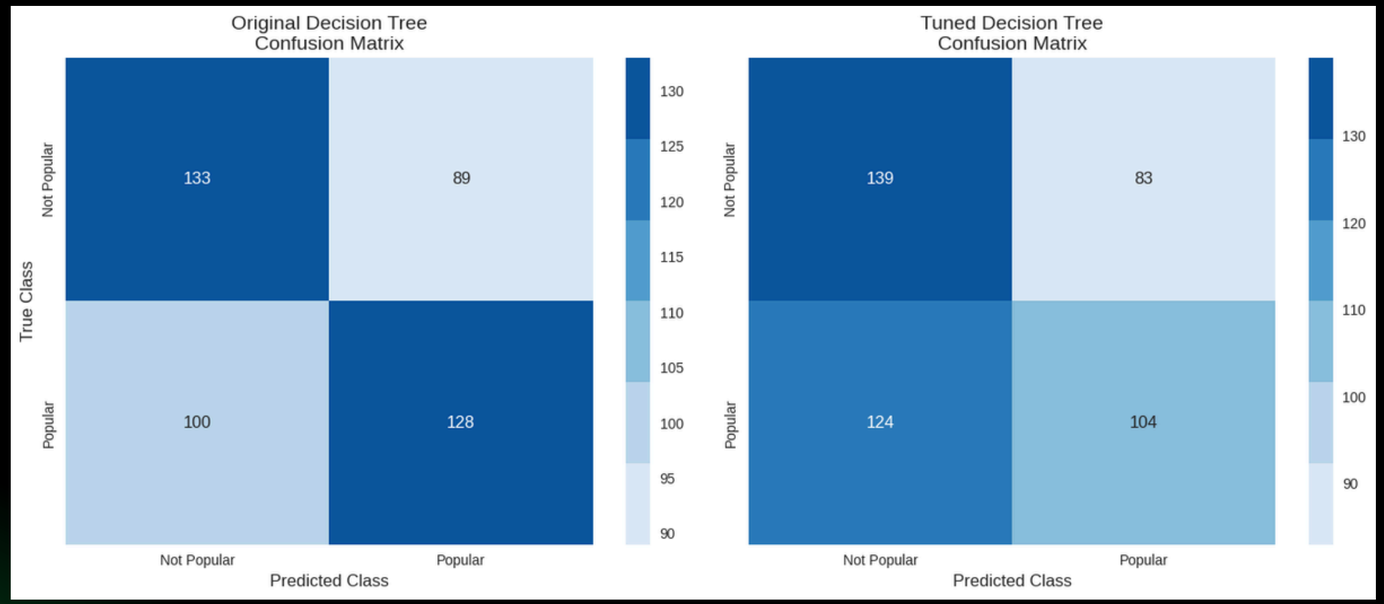
The VP of sales has indicated that they would like to maximize the monetary value of the predictions (i.e., maximize revenue from correct predictions less the cost of prediction errors). The revenue and costs are tied to the confusion matrix. The revenue for correctly predicting a popular song is $1,000. The cost for incorrectly predicting a song is popular, when in fact it is not, is $700. The cost for incorrectly predicting a song is not popular, when in fact it is, is $900. There is no cost or revenue for correctly predicting a song is non-popular.

Methodology followed:

- Used the confusion matrix for each of the models
  - Decision Tree
  - K-Nearest Neighbor
  - Logistic Regression

- Input the revenue and costs tied to the confusion matrix
- Compared the best model selected in Q1 with the best model selected in Q2.

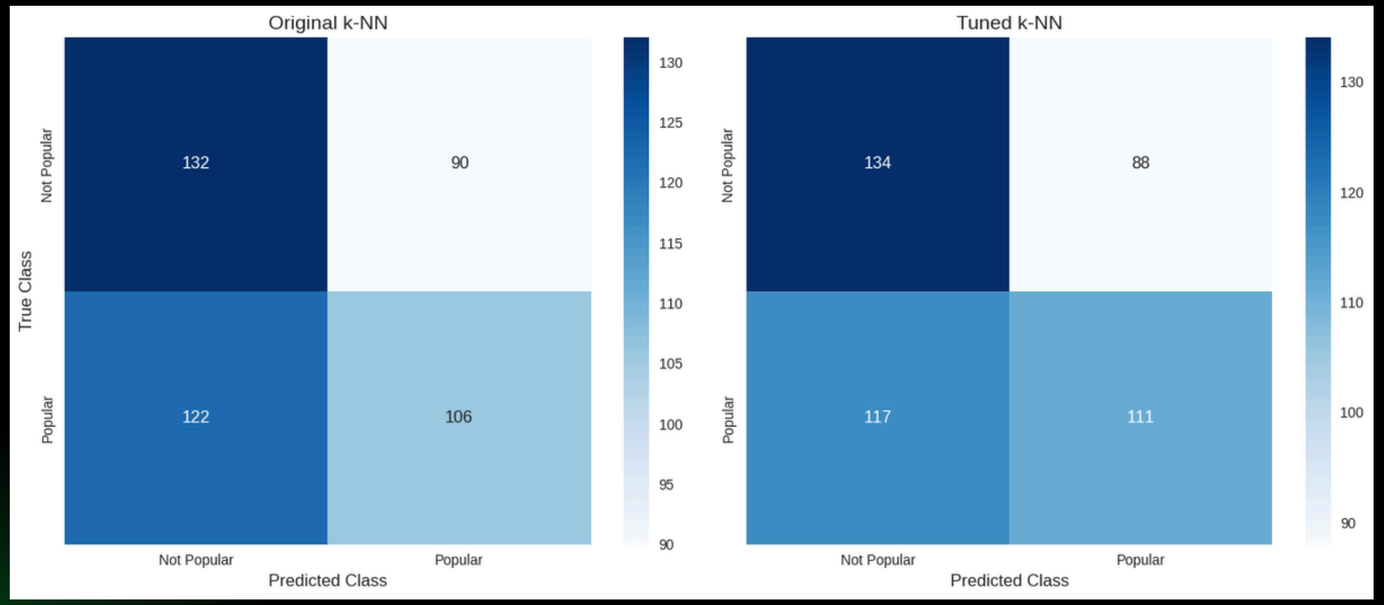
Decision Tree

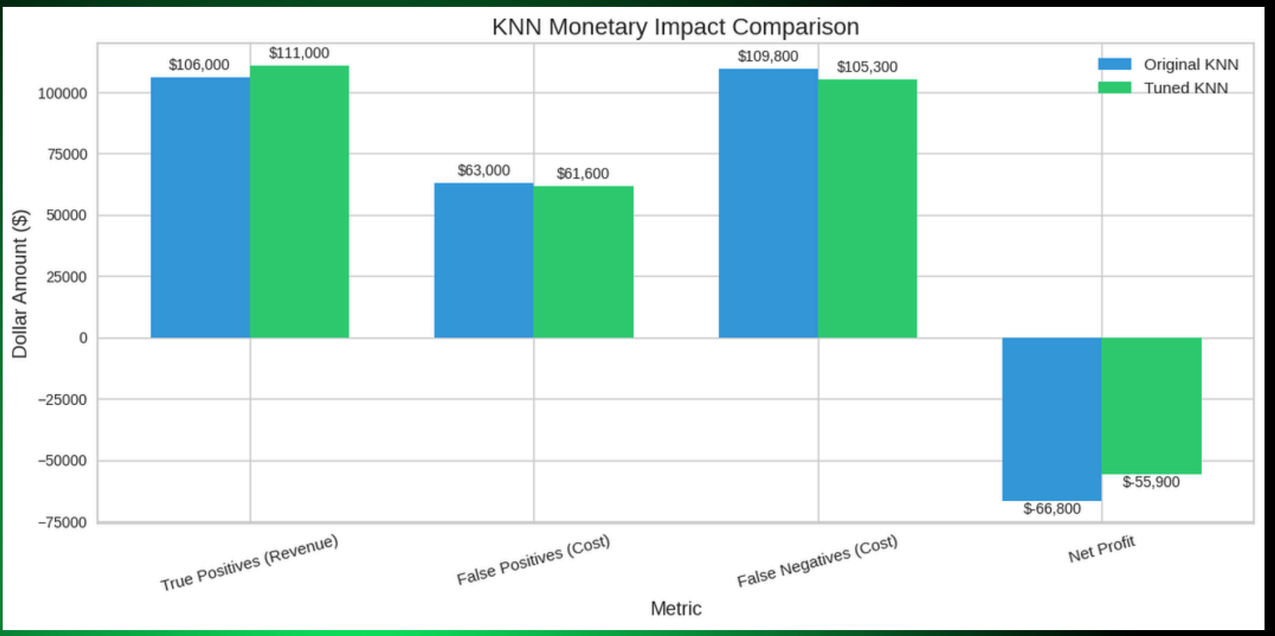


💰 Monetary Impact Analysis (Q2):

	Metric	Original Decision Tree	Tuned Decision Tree
0	True Positives (Revenue)	$128,000	$104,000
1	False Positives (Cost)	$62,300	$58,100
2	False Negatives (Cost)	$90,000	$111,600
3	Net Profit	$-24,300	$-65,700

K-Nearest Neighbor

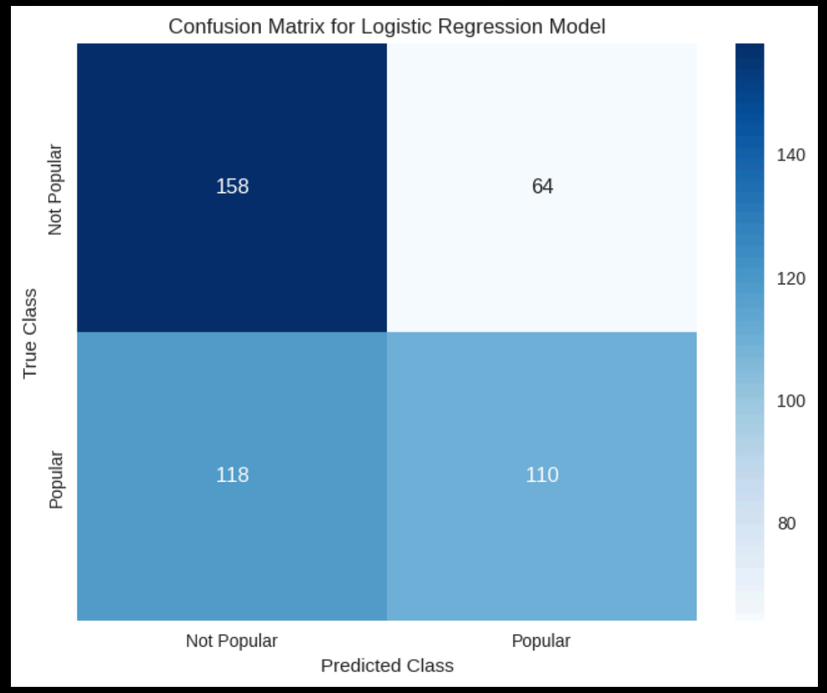


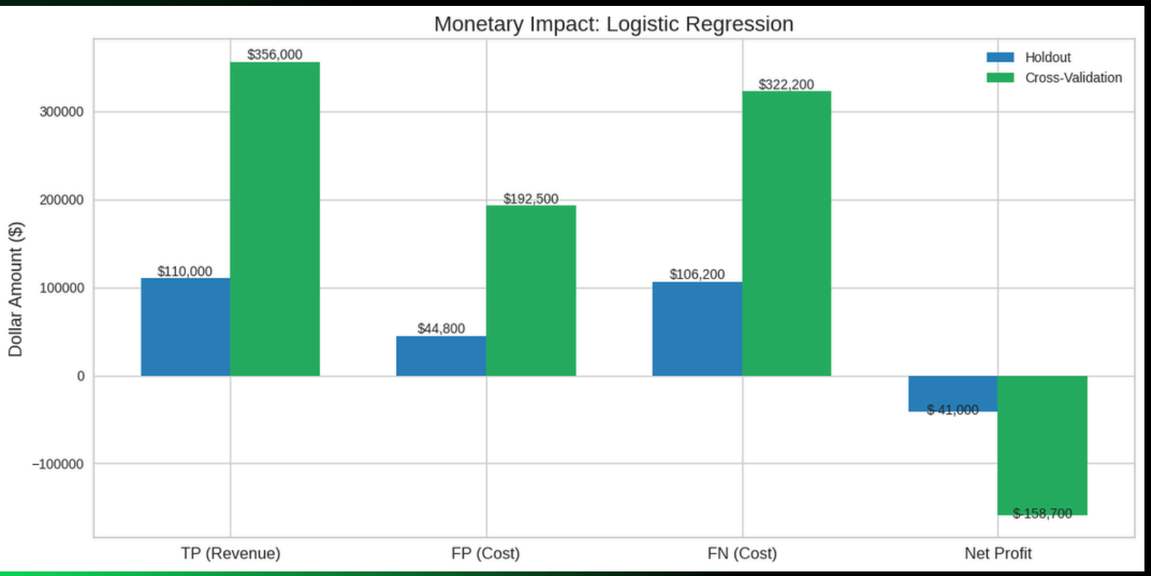


💰 Monetary Impact Analysis - KNN (Q2):

	Metric	Original KNN	Tuned KNN
0	True Positives (Revenue)	$106,000	$111,000
1	False Positives (Cost)	$63,000	$61,600
2	False Negatives (Cost)	$109,800	$105,300
3	Net Profit	$-66,800	$-55,900

Logistic Regression





💰 Logistic Regression: Monetary Impact Analysis

	Metric	Logistic Regression (Holdout)	Logistic Regression (CV)
0	True Positives (Revenue)	$110,000	$356,000
1	False Positives (Cost)	$44,800	$192,500
2	False Negatives (Cost)	$106,200	$322,200
3	Net Profit	$-41,000	$-158,700

Evaluation

Logistic Regression was selected based on its highest test accuracy (0.5959), outperforming both Decision Tree and k-NN models in generalization performance.

Under the profit-based evaluation in Q2, the original decision tree model yielded the least financial loss (-$24,300), making it the most viable option when maximizing monetary return is the goal.

Criterion	Logistic Regression (Q1)	Original Decision Tree (Q2)
Selection Goal	Maximize Predictive Accuracy	Maximize Monetary Profit
Test Accuracy	59.60%	58.00%
Net Profit	–$41,000	–$24,300
True Positives ($)	$110,000	$128,000
False Negatives ($)	$106,200	$90,000

Managerial Insights

While accuracy-driven models (like Logistic Regression) performed better in predicting song popularity, the **monetary impact analysis revealed** a different priority: minimizing costly misclassifications is essential when revenue and operational efficiency are on the line. The original Decision Tree, though less accurate, produced the **least financial loss (–$24,300)**, primarily by identifying more profitable true positives.

Trade-Offs Identified:

- **Accuracy vs Revenue**
- **Precision vs Recall**

Recommendation :

We recommend a **dual-model strategy** to optimize both user engagement and revenue in Spotify's recommendation system.

- Adopt Logistic Regression for personalized playlists
- Deploy Decision Tree for revenue-focused promotions
- Implement a hybrid strategy

Question 3

The company is also interested in understanding:

- The effects of valence of a song in predicting the song's success and whether the effect differs across different types of music.
- Which features or combinations of features are frequently associated with popularity of songs. This analysis would help improve their playlist recommendations by identifying similar songs based on these patterns.

Methodology followed:

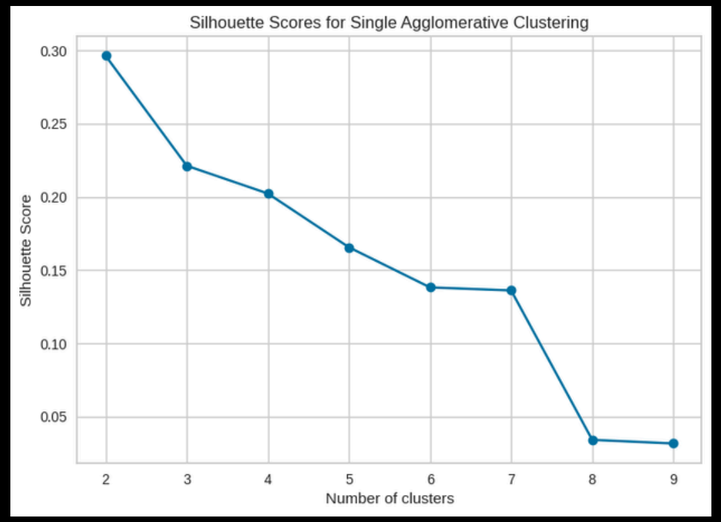
- Used the logistic regression model for identifying the effects of valence of a song to predict it's success
- Used Agglomarative clsutering to cluster all the songs in the datasets by danceability and energy
- Examine the cluster characteristics to identify the best performing cluster
- Developed a Association rule model, to identify combinations of features are frequently associated with popularity of songs.

Effects of Valence

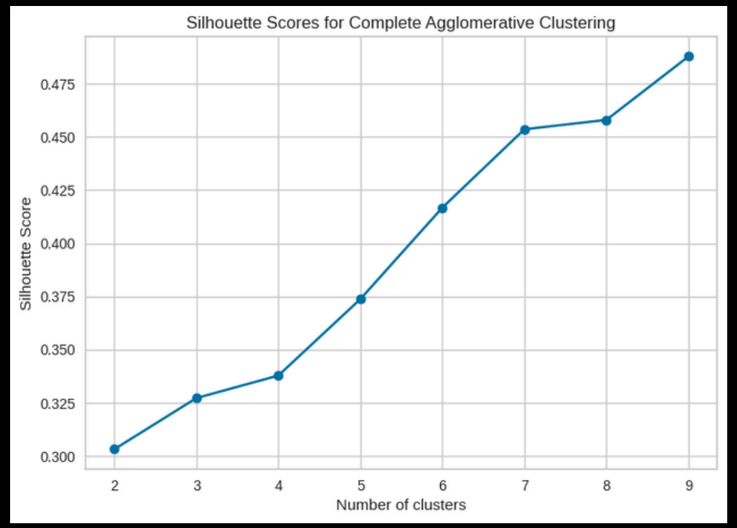
```
Optimization terminated successfully.
         Current function value: 0.656953
         Iterations 5
                         Results: Logit
================================================================
Model:              Logit            Method:           MLE
Dependent Variable: popularity_label Pseudo R-squared: 0.048
Date:               2025-04-24 17:15 AIC:              1419.6011
No. Observations:   1050             BIC:              1518.7320
Df Model:           19               Log-Likelihood:   -689.80
Df Residuals:       1030             LL-Null:          -724.90
Converged:          1.0000           LLR p-value:      8.4927e-08
No. Iterations:     5.0000           Scale:            1.0000
----------------------------------------------------------------
                  Coef.  Std.Err.    z    P>|z|   [0.025  0.975]
----------------------------------------------------------------
song_name_len    -0.1288   0.0660 -1.9511 0.0511 -0.2582  0.0006
duration_ms       0.0694   0.0679  1.0210 0.3073 -0.0638  0.2025
explicit         -0.1162   0.0777 -1.4956 0.1347 -0.2686  0.0361
danceability     -0.0123   0.0831 -0.1486 0.8819 -0.1752  0.1505
energy           -0.1936   0.0990 -1.9547 0.0506 -0.3877  0.0005
loudness          0.3244   0.0858  3.7807 0.0002  0.1562  0.4926
mode             -0.0963   0.0650 -1.4813 0.1385 -0.2237  0.0311
speechiness      -0.0798   0.0758 -1.0526 0.2925 -0.2284  0.0688
acousticness      0.0856   0.0733  1.1690 0.2424 -0.0579  0.2292
instrumentalness -0.0706   0.0761 -0.9282 0.3533 -0.2197  0.0785
liveness         -0.0724   0.0660 -1.0964 0.2729 -0.2018  0.0570
valence           0.0202   0.0822  0.2463 0.8054 -0.1408  0.1813
tempo             0.1841   0.0673  2.7367 0.0062  0.0523  0.3160
pop              -0.1047   0.0706 -1.4829 0.1381 -0.2431  0.0337
rock              0.2638   0.0763  3.4589 0.0005  0.1143  0.4133
hiphop            0.2484   0.0812  3.0597 0.0022  0.0893  0.4076
dance            -0.1410   0.0707 -1.9949 0.0461 -0.2795 -0.0025
folk             -0.0270   0.0588 -0.4593 0.6460 -0.1423  0.0883
rnb              -0.1785   0.0724 -2.4667 0.0136 -0.3203 -0.0367
latin             0.0014   0.0664  0.0213 0.9830 -0.1287  0.1315
================================================================
```

	Feature	Coefficient	Odds Ratio
5	loudness	0.3200	1.3771
14	rock	0.2603	1.2974
15	hiphop	0.2464	1.2793
12	tempo	0.1838	1.2017
8	acousticness	0.0838	1.0874
1	duration_ms	0.0687	1.0711
11	valence	0.0257	1.0260

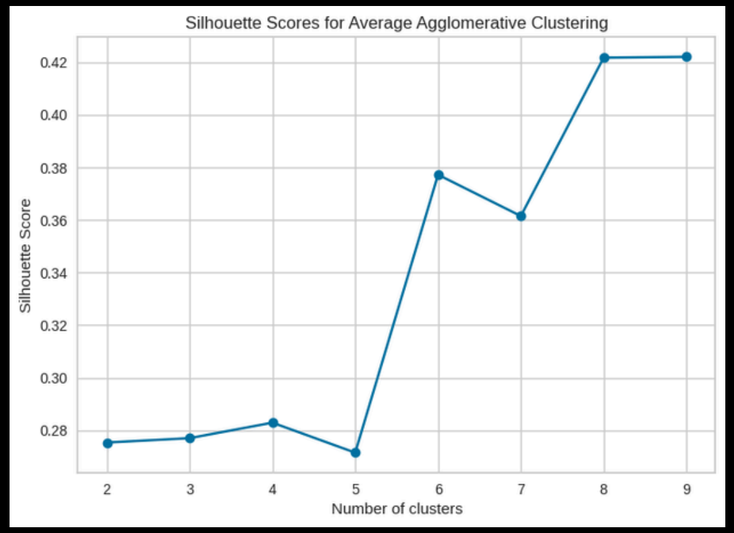
Agglomerative Clustering



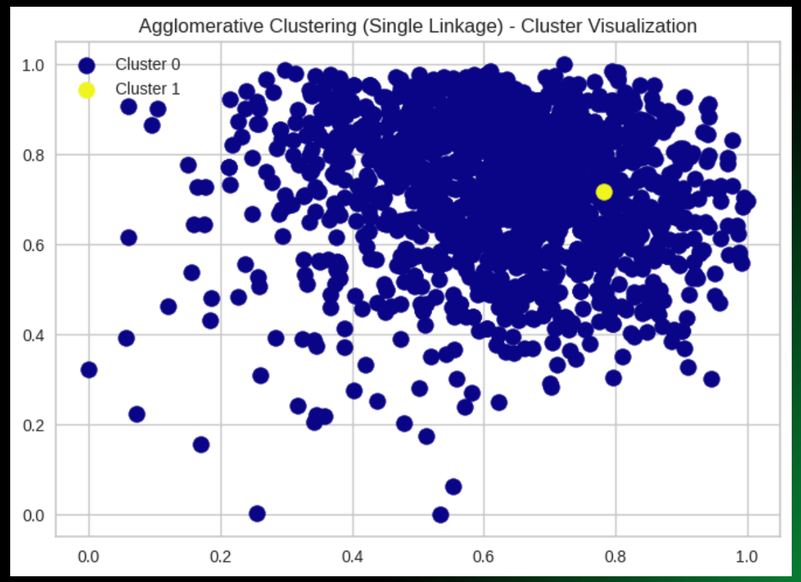
Average silhouette score for Single Linkage: 0.29629

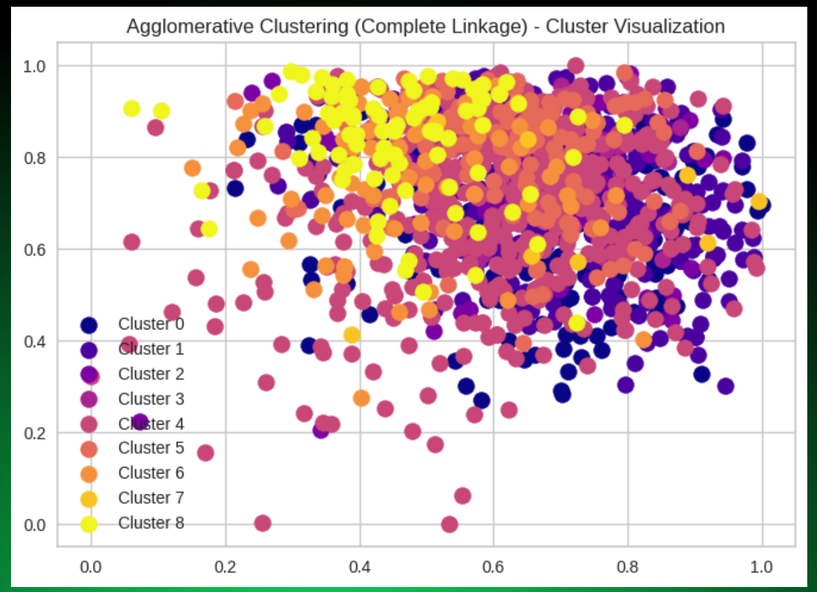


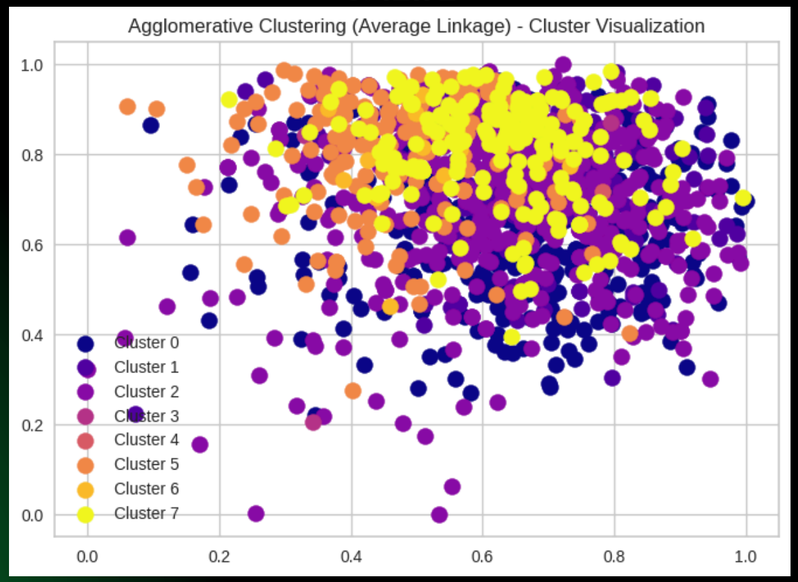
Average silhouette score for Complete linkage: 0.48777



Average silhouette score for Average linkage: 0.42172







Agglomerative Clustering Profiles

	song_name_len	duration_ms	explicit	danceability	energy	loudness	mode	speechiness	acousticness	instrumentalness	liveness	valence	tempo	pop	rock	hiphop	dance	folk	rnb	latin	popularity_label	Count	Percentage
Cluster 0	17.56	234691.69	0.24	0.66	0.74	-5.43	0.56	0.10	0.12	0.02	0.19	0.57	119.89	0.82	0.13	0.39	0.16	0.01	0.26	0.03	0.48	1499	99.93
Cluster 1	7.00	199040.00	0.00	0.79	0.73	-5.12	0.00	0.03	0.00	0.20	0.09	0.92	105.99	1.00	1.00	0.00	0.00	1.00	1.00	0.00	0.00	1	0.07
Overall	17.55	234667.92	0.24	0.66	0.74	-5.43	0.56	0.10	0.12	0.02	0.19	0.57	119.88	0.82	0.13	0.39	0.16	0.01	0.26	0.03	0.48	1500	100.00

Cluster 1 characteristics - Single Linkage

	song_name_len	duration_ms	explicit	danceability	energy	loudness	mode	speechiness	acousticness	instrumentalness	liveness	valence	tempo	pop	rock	hiphop	dance	folk	rnb	latin	popularity_label	Count	Percentage
Cluster 2	17.28	233949.07	0.30	0.67	0.73	-5.48	0.59	0.11	0.13	0.01	0.19	0.58	119.69	0.99	0.02	0.40	0.00	0.00	0.00	0.06	0.50	574	38.27
Cluster 0	16.77	242546.40	0.25	0.69	0.67	-5.83	0.50	0.11	0.16	0.01	0.16	0.58	114.42	0.97	0.00	0.57	0.01	0.00	1.00	0.01	0.40	391	26.07
Cluster 7	21.33	216659.65	0.06	0.66	0.81	-4.76	0.55	0.07	0.07	0.05	0.20	0.58	125.77	0.99	0.08	0.23	1.00	0.00	0.03	0.00	0.40	195	13.00
Cluster 5	15.07	230351.27	0.06	0.52	0.81	-5.05	0.63	0.05	0.05	0.03	0.21	0.51	127.98	0.46	1.00	0.00	0.00	0.00	0.00	0.01	0.65	171	11.40
Cluster 1	19.01	246022.17	0.47	0.69	0.76	-5.47	0.56	0.14	0.11	0.04	0.20	0.58	118.80	0.00	0.00	0.55	0.25	0.00	0.00	0.00	0.46	153	10.20
Cluster 6	14.42	224269.75	0.08	0.54	0.79	-5.93	0.83	0.05	0.09	0.00	0.21	0.57	117.13	1.00	0.25	0.00	0.00	1.00	0.00	0.00	0.50	12	0.80
Cluster 3	9.67	236368.33	0.00	0.54	0.66	-7.77	0.67	0.04	0.33	0.22	0.26	0.59	96.66	0.00	0.67	0.00	0.00	1.00	0.00	0.00	0.67	3	0.20
Cluster 4	7.00	199040.00	0.00	0.79	0.73	-5.12	0.00	0.03	0.00	0.20	0.09	0.92	105.99	1.00	1.00	0.00	0.00	1.00	1.00	0.00	0.00	1	0.07
Overall	17.55	234667.92	0.24	0.66	0.74	-5.43	0.56	0.10	0.12	0.02	0.19	0.57	119.88	0.82	0.13	0.39	0.16	0.01	0.26	0.03	0.48	1500	100.00

Cluster 2 characteristics - Average Linkage

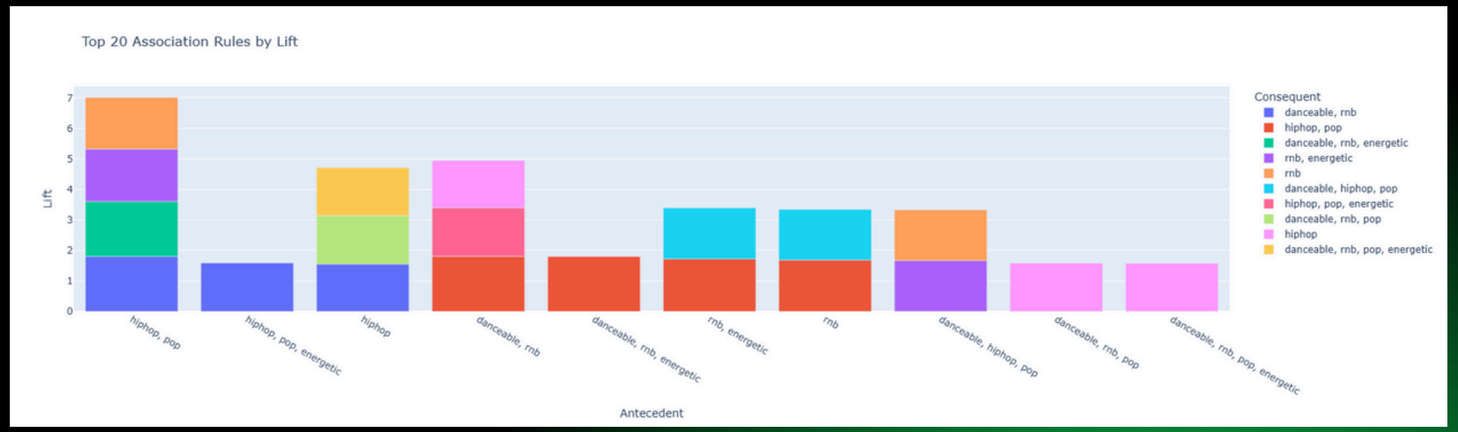
	song_name_len	duration_ms	explicit	danceability	energy	loudness	mode	speechiness	acousticness	instrumentalness	liveness	valence	tempo	pop	rock	hiphop	dance	folk	rnb	latin	popularity_label	Count	Percentage
Cluster 4	16.37	233343.56	0.10	0.65	0.71	-5.61	0.57	0.07	0.16	0.01	0.17	0.56	119.12	1.00	0.00	0.00	0.00	0.02	0.32	0.00	0.46	495	33.00
Cluster 1	19.04	241455.90	0.61	0.71	0.74	-5.17	0.56	0.17	0.10	0.01	0.21	0.59	117.18	0.76	0.03	1.00	0.15	0.00	0.00	0.00	0.53	343	22.87
Cluster 0	16.95	242879.69	0.36	0.71	0.67	-5.92	0.47	0.13	0.13	0.00	0.16	0.60	114.58	0.94	0.00	0.96	0.00	0.00	1.00	0.01	0.43	229	15.27
Cluster 5	19.89	215836.23	0.03	0.65	0.81	-5.02	0.54	0.07	0.07	0.06	0.18	0.58	126.30	0.99	0.10	0.00	1.00	0.00	0.00	0.00	0.34	145	9.67
Cluster 8	15.54	226249.12	0.11	0.52	0.85	-4.64	0.57	0.06	0.03	0.04	0.21	0.51	128.43	0.00	1.00	0.00	0.00	0.02	0.00	0.00	0.77	92	6.13
Cluster 6	13.76	235061.70	0.02	0.53	0.77	-5.54	0.70	0.04	0.07	0.01	0.21	0.50	126.63	1.00	1.00	0.00	0.00	0.05	0.01	0.00	0.53	83	5.53
Cluster 2	20.24	236316.53	0.10	0.66	0.76	-5.80	0.64	0.05	0.13	0.07	0.18	0.57	123.65	0.00	0.00	0.00	0.46	0.01	0.00	0.00	0.30	70	4.67
Cluster 3	23.83	230053.14	0.03	0.71	0.81	-5.09	0.69	0.08	0.10	0.00	0.20	0.72	121.53	0.86	0.06	0.37	0.00	0.00	0.00	1.00	0.46	35	2.33
Cluster 7	13.75	230336.25	0.25	0.74	0.69	-6.13	0.62	0.14	0.14	0.00	0.16	0.64	110.37	1.00	0.00	0.38	1.00	0.00	1.00	0.00	0.38	8	0.53
Overall	17.55	234667.92	0.24	0.66	0.74	-5.43	0.56	0.10	0.12	0.02	0.19	0.57	119.88	0.82	0.13	0.39	0.16	0.01	0.26	0.03	0.48	1500	100.00

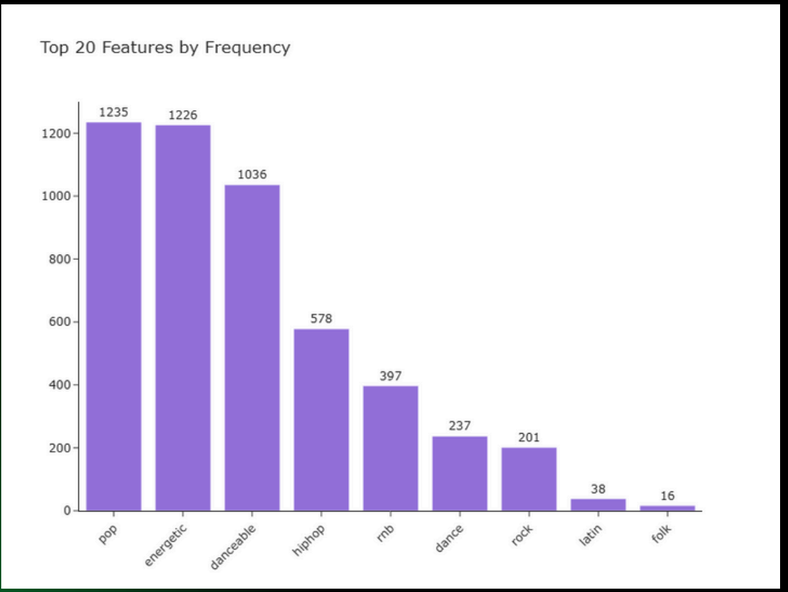
Cluster 3 characteristics - Complete Linkage

Association Rule

	Antecedent	Consequent	Support	Confidence	Lift
0	hiphop, pop	danceable, rnb	0.1213	0.3707	1.8052
1	danceable, rnb	hiphop, pop	0.1213	0.5909	1.8052
2	danceable, rnb, energetic	hiphop, pop	0.0847	0.5880	1.7962
3	hiphop, pop	danceable, rnb, energetic	0.0847	0.2587	1.7962
4	rnb, energetic	hiphop, pop	0.1033	0.5636	1.7219
5	hiphop, pop	rnb, energetic	0.1033	0.3157	1.7219
6	hiphop, pop	rnb	0.1460	0.4460	1.6852
7	rnb	hiphop, pop	0.1460	0.5516	1.6852
8	danceable, hiphop, pop	rnb, energetic	0.0847	0.3060	1.6692
9	rnb, energetic	danceable, hiphop, pop	0.0847	0.4618	1.6692

	Antecedent	Consequent	Support	Confidence	Lift
0	danceable	hiphop	0.3233	0.4681	1.2149
1	hiphop	danceable	0.3233	0.8391	1.2149
2	hiphop, pop	danceable	0.2767	0.8452	1.2238
3	danceable	hiphop, pop	0.2767	0.4006	1.2238
4	danceable, pop	hiphop	0.2767	0.4716	1.2239
5	hiphop	danceable, pop	0.2767	0.7180	1.2239
6	danceable	hiphop, pop, energetic	0.2160	0.3127	1.2059
7	hiphop, pop, energetic	danceable	0.2160	0.8329	1.2059
8	danceable, pop	rnb	0.1987	0.3386	1.2795
9	rnb	danceable, pop	0.1987	0.7506	1.2795





Managerial Insights

- Moderate Valence with High Loudness Performs Better
- Genre Combinations Drive Engagement
- Optimize Playlist Diversity Using Feature Clusters

Recommendation:

Prioritize High-Performing Sound Profiles in Playlists

Maximize Reach with Multi-Genre Popular Crossovers

Streamline User Experience by Filtering Low-Engagement Tracks

Enhance Personalization Through Cluster-Based Playlist Themes

Use Valence Strategically to Match Listener Preferences

Thank You