Music Analytics

Report by Brian Pan

Overview

In this report, we are analyzing music data for five popular contemporary artists in the melodic electronic music genre. The data is collected from Spotify and is then manipulated for further analysis. In our analysis, we will look at a few popular artists, namely Dabin, Gryffin, Illenium, Sabai, and Seven Lions. While these artists do not reflect the entirety of the subgenre of electronic music, it is a good step in reviewing the musical tastes of the subgenre's audience and see how to further capture and expand this audience.

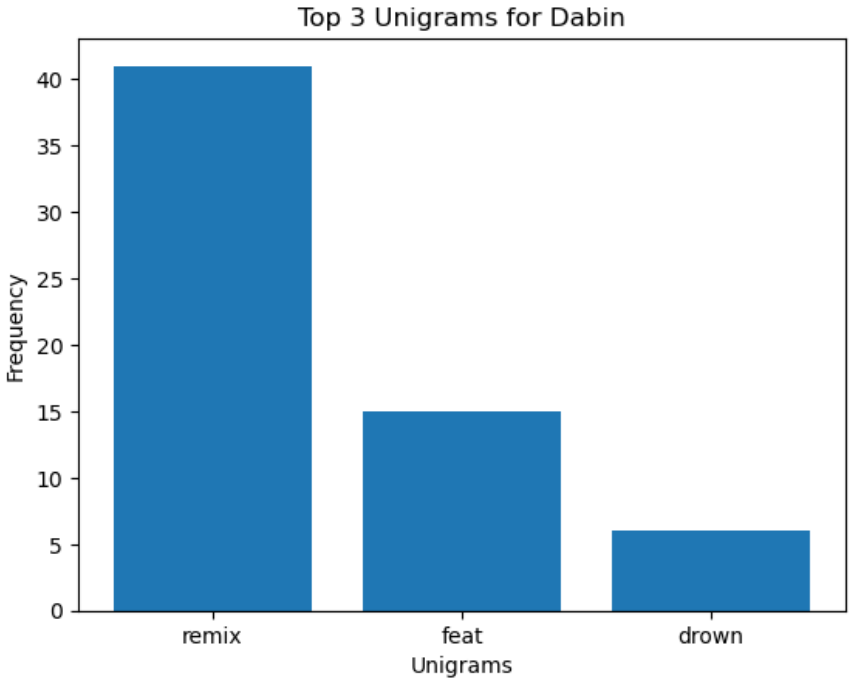
Below is the general overview of the report:

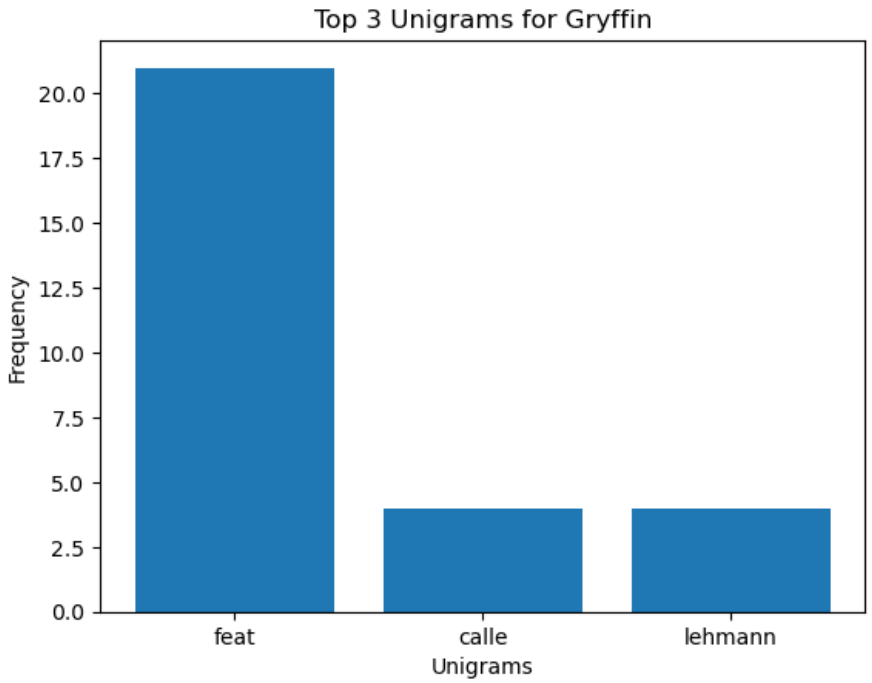
- Descriptive Analytics of Song Names
    - Unigram
    - Bigram
    - Trigram
- Regression Analysis of Sound Characteristics
    - Correlation
    - VIF
    - Lasso
    - Ridge
    - Random Forest

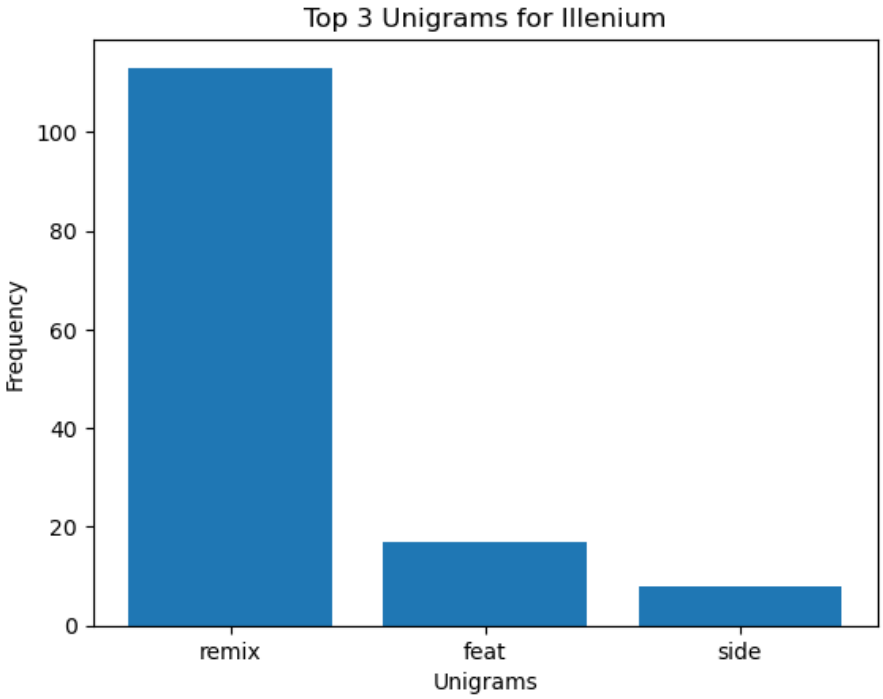
- General Overview of Cover Art
    - Cover Art Trends
    - Designs by Popularity

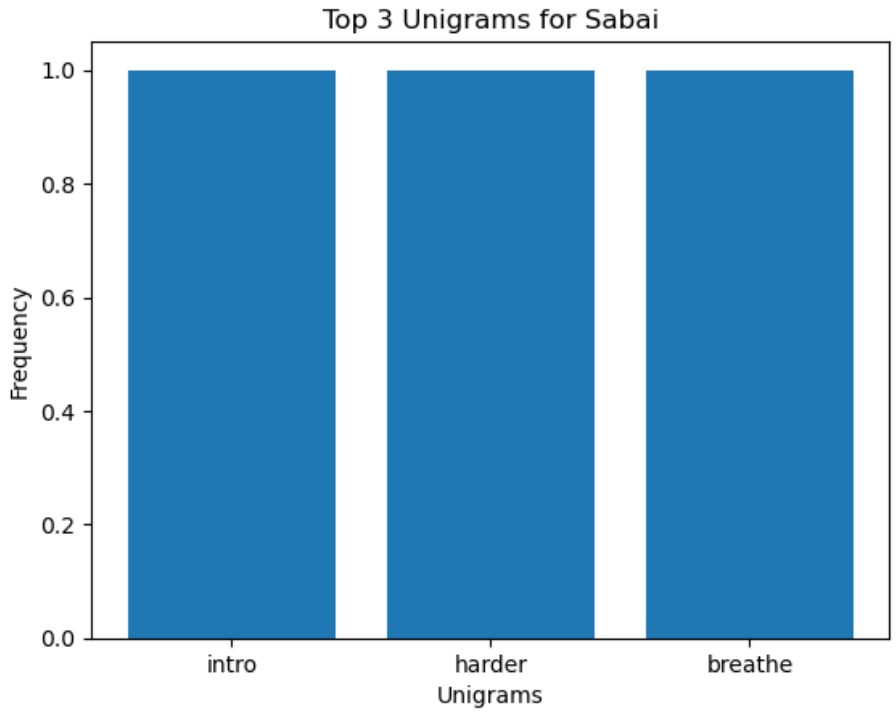
Descriptive Analytics of Song Names

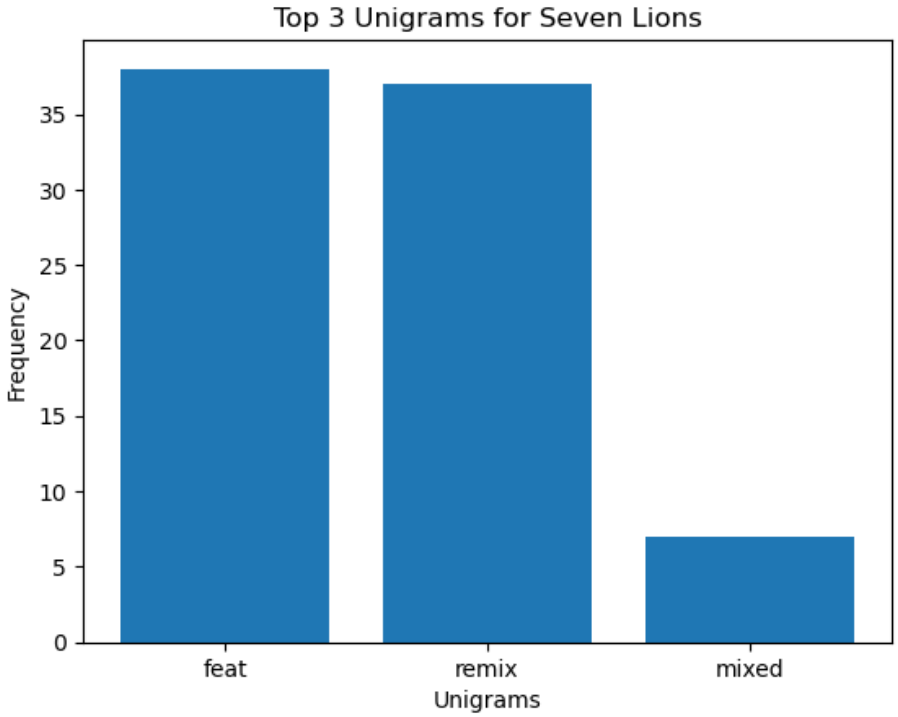
While song names often reflect a catchy tune or clue into the sentiments of the melody, like book titles, the name of the song have implicit sentimental value for the listener. The following charts were created by extracting song names from across all of the artists' albums and seeing which were the most popular words that appeared. Stop words and unnecessary punctuation such as 'a', 'the', '-', etc were removed to provide more meaningful insight. These select key words are then aggregated and tallied into the database, where the first (or top) three are then visualized below.




Top 3 Unigrams for Gryffin
Frequency
20.0
17.5
15.0
12.5
10.0
7.5
5.0
2.5
0.0
feat
calle
lehmann
Unigrams


Top 3 Unigrams for Illenium
Frequency
100
80
60
40
20
0
remix
feat
side
Unigrams





A general review of all the charts, we can see a lot of trends. While these trends do not capture any common sentiments, it does offer a new insight previously unanticipated. Song names often include information providing extensions. This may be indicative of a

remix, a cover, or even a collaboration. This is a common way that artists may attempt to expand their target audience base by relying on other artists and capitalizing on the intersection of fan bases. A collaboration can, for instance, link two previously unrelated fan bases together and fans previously unaware of the artist are now exposed.

A summary of the findings are as follows:

	Dabin	Gryffin	Illenium	Sabai	Seven Lions
'feat'	Yes	Yes	Yes		Yes
'remix'	Yes		Yes		Yes
'mixed'					Yes

From the summary chart, we can easily see that Sabai does not do much collaboration or remixes. This may be because of several things. More experienced artists, or artists with more releases, are more likely to collaborate, mix, or remix songs as opposed to newer artists who are still building their fan base. One stark example is Seven Lions, who has all three.

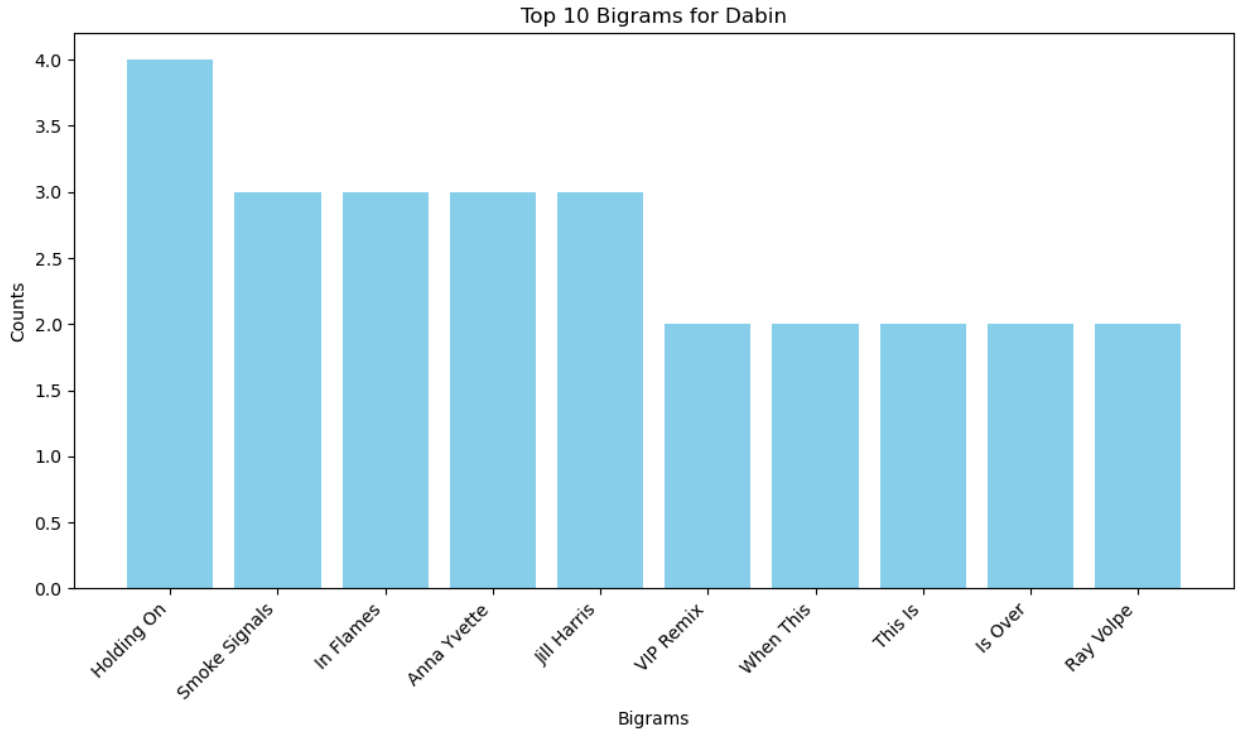
- 'feat' indicates a collaboration with another artist. Reference to the Unigram chart for Seven Lions indicates that more than 35 titles has included the key term 'feat.' This suggests that Seven Lions is a big collaborator.
- 'remix' are re-interpretations of songs. This could be a cover of another artist or a personal song. Having a similar high count around 35 titles could suggest that Seven Lions is exploring ways to diversify or expand audience by experimenting with different music tastes.
- 'mixed' serves as a distinction from 'remix.' While this may be interpreted as several ways, it usually means the complimentary meaning of 'remix.' If 'remix' suggests a cover of another artist, 'mix' might suggest a re-interpretation of an original song and vice versa. Seven Lions having 'mix' as a substantial re-occurring instance is more indicative of how prolific Seven Lions is and how he may be well entrenched in the industry.

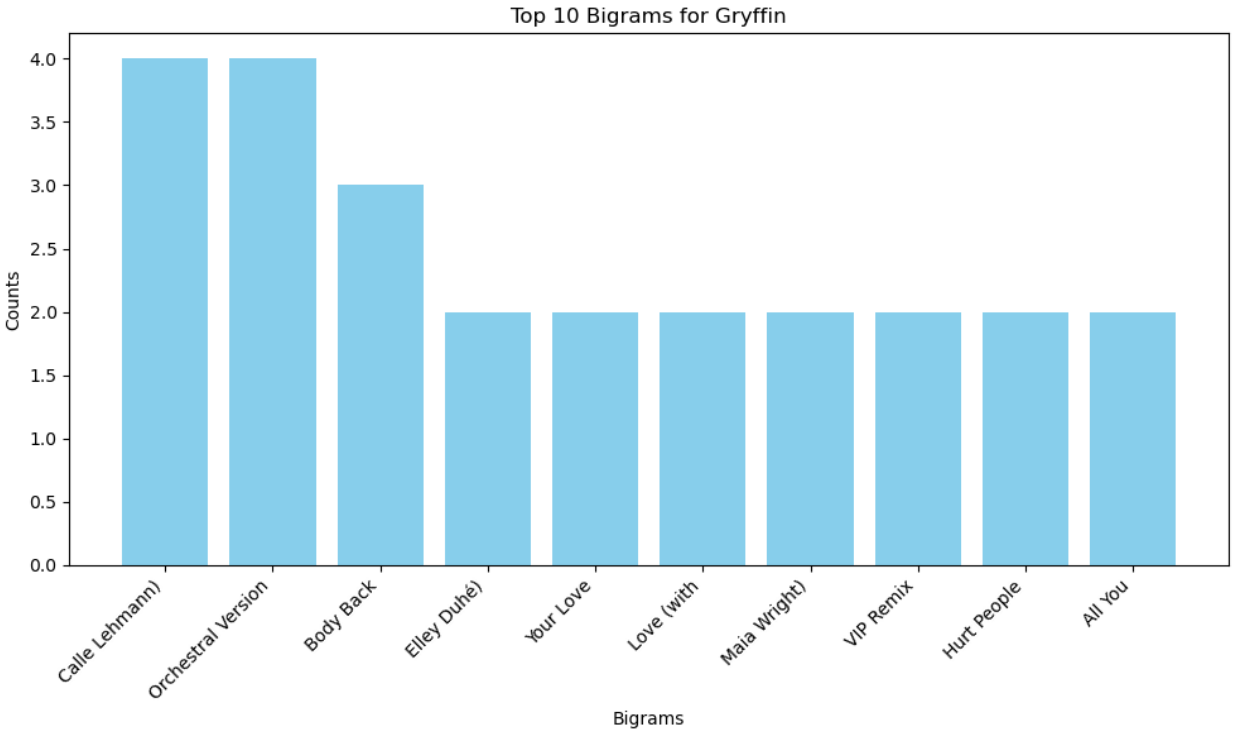
Another overarching theme we can glean off this is that across the five artists, four have 'feat' and three have 'remix'. Collaboration is a heavy factor in this industry, and it may even be expected for artists to collaborate with one another at one point.

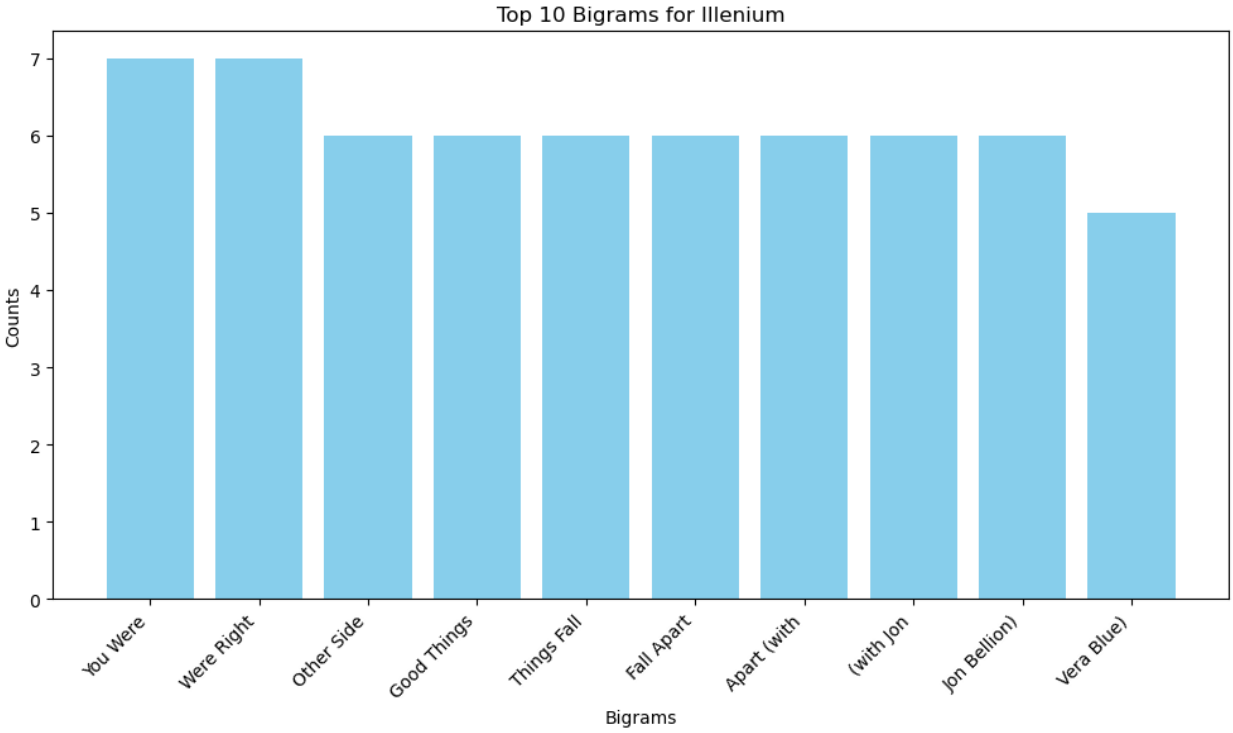
Bigrams

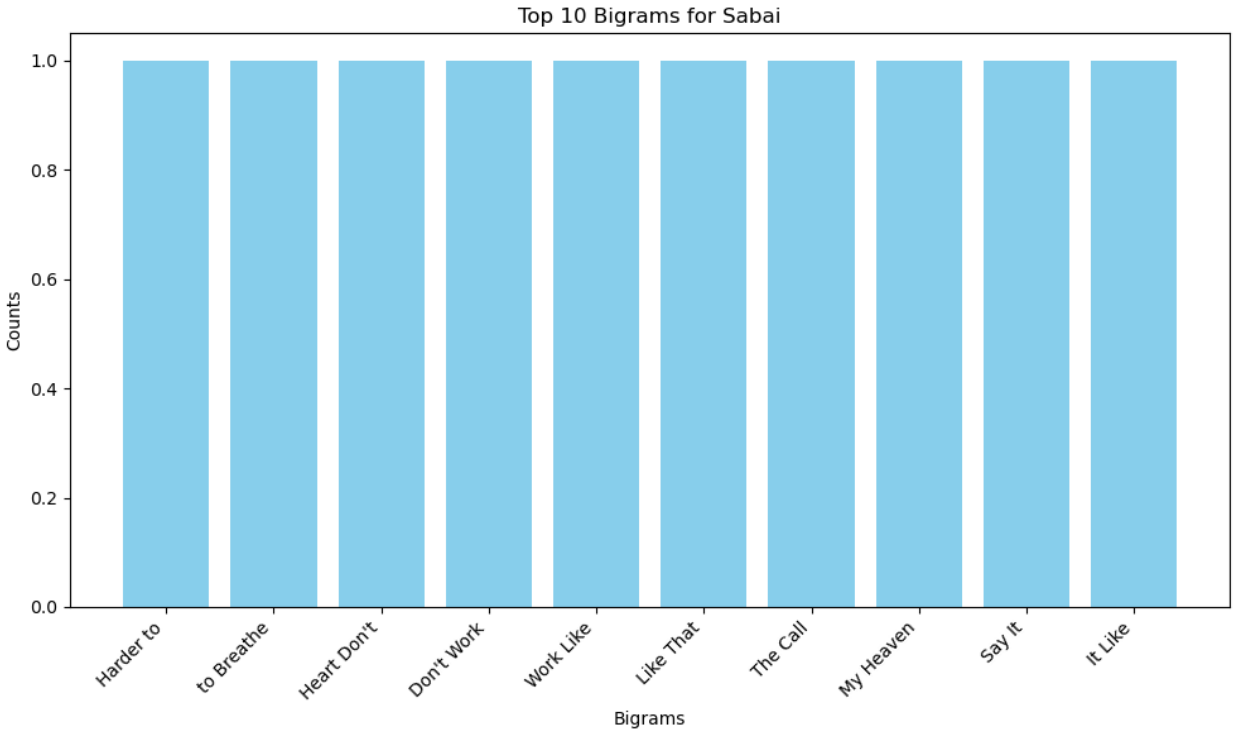
Sometimes single words (unigrams) are not sufficient in understanding the details of the data. To delve in further, we will look at bigrams, which take clusters of two-words together. The sentence 'This is Brian' would thus be extracted as two items: 'This is' and 'is Brian.' From a language perspective, this will allow us to better capture word clusters like "New Age" (synonym of contemporary) which separated loses its intended meaning.

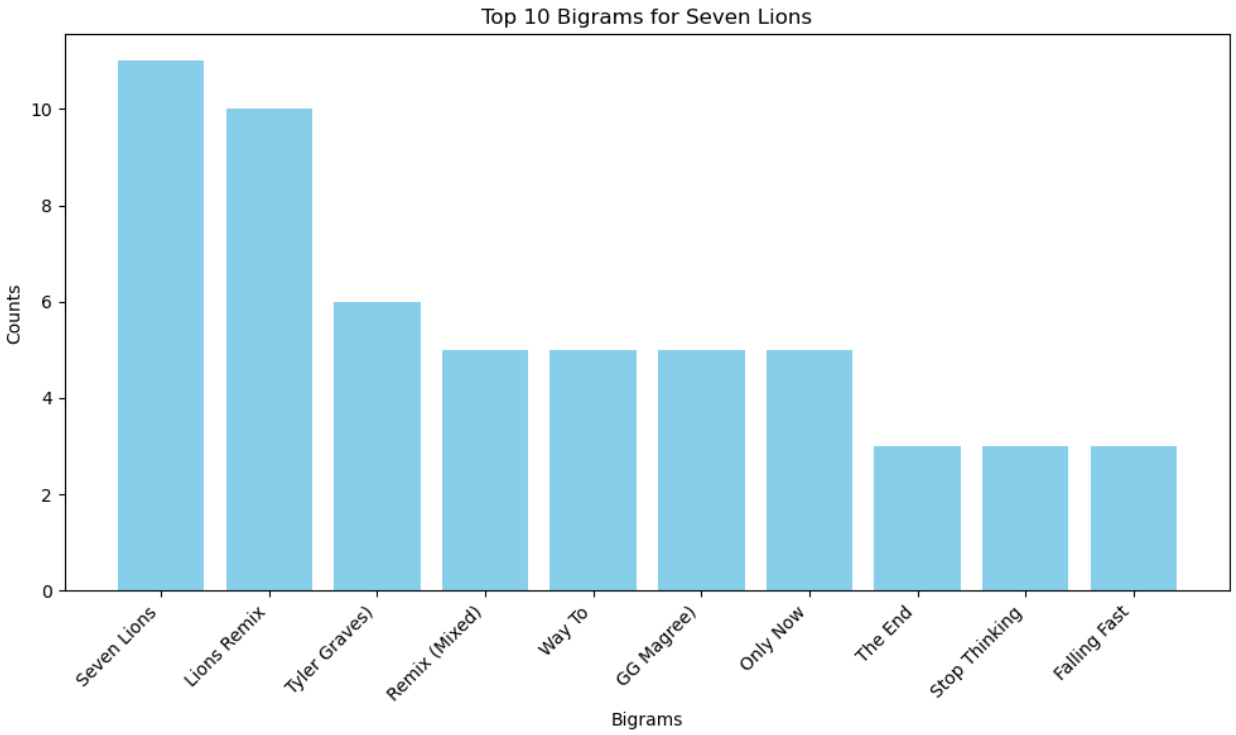
To derive our two-word clusters, we first use the scraped data from Spotify and iterate over the rows to add the sequential titles to a new list. Afterwards, we run a bigram algorithm that separates the words into bigrams. We then filter out unnecessary words and punctuation and iterate this list into a dictionary with keys as the bigrams and value as the count values. This dictionary is sorted and then plotted for the following results.




Top 10 Bigrams for Gryffin
Counts
Bigrams
4.0
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
Calle Lehmann)
Orchestral Version
Body Back
Elley Duhé)
Your Love
Love (with
Maia Wright)
VIP Remix
Hurt People
All You


Top 10 Bigrams for Illenium
Counts
Bigrams
7
6
5
4
3
2
1
0
You Were
Were Right
Other Side
Good Things
Things Fall
Fall Apart
Apart (with
(with Jon
Jon Bellion)
Vera Blue)





Some interesting findings come up with the bigram analysis. Collaboration artists, especially frequent collaborations, are now appear in their full name and are easier to spot.

Popular songs and remixes are also prevalent and appear in their full glory. If nothing noteworthy was found, the bigrams in the bar chart are static and have no change with all same/similar count values (see Sabai). Below is a summary of the bigram charts.

	Dabin	Gryffin	Illenium	Sabai	Seven Lions
Remixed Titles	Smoke Signal Holding On In Flame	Body back	You Were Right Other Side		Only Now
Collaborator	Jill Harris	Calle Lehmann Elley Duhe	Jon Bellion Vera Blue		Tyler Graves GG Magree
Versions		Orchestral			Seven Lions Remix

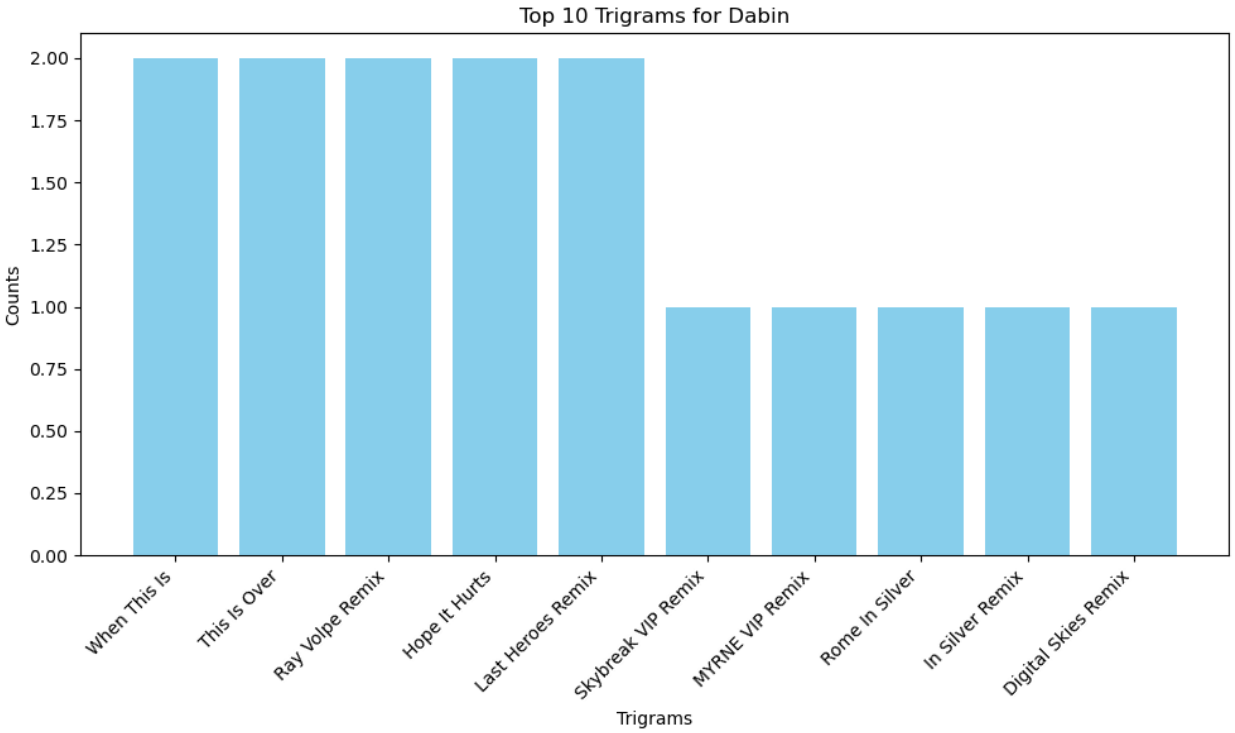
Some noteworthy distinctions below:

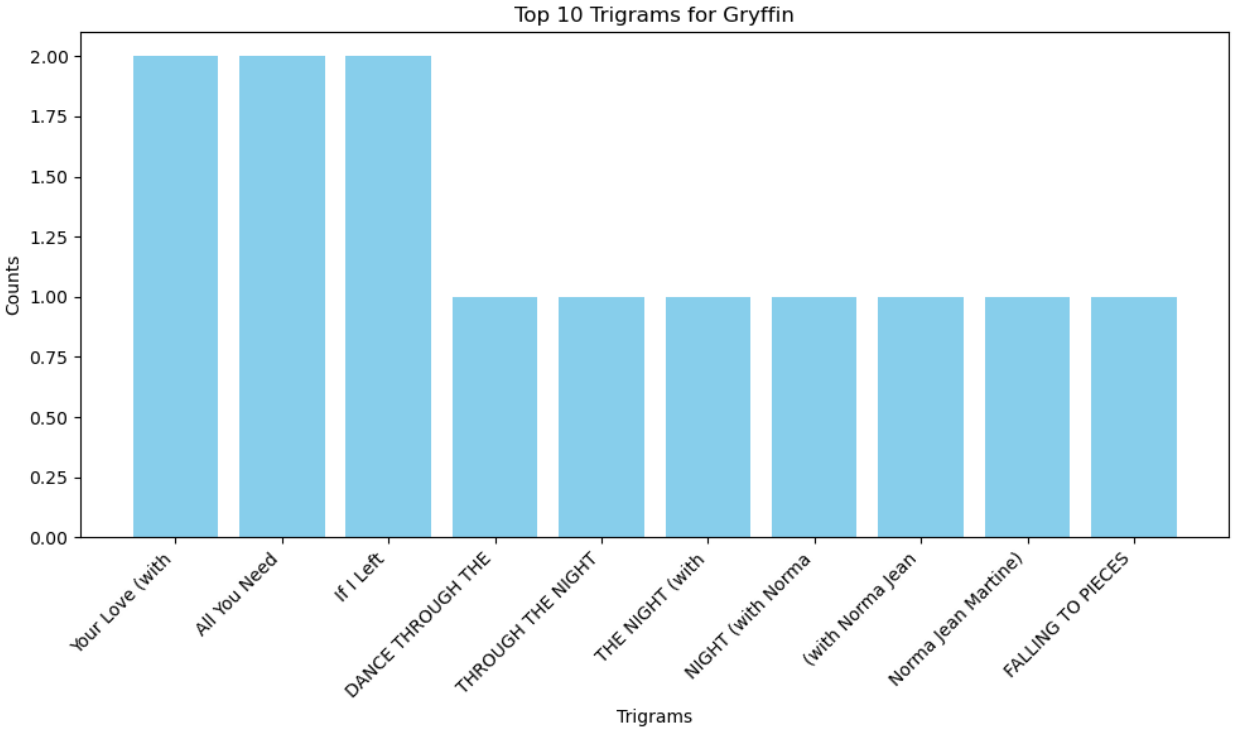
- Gryffin is the only one with orchestral version
- Seven Lions frequently remixes for other artists
- All except Sabai have had collaborations with other artists
- Almost all have had multiple versions of hit songs

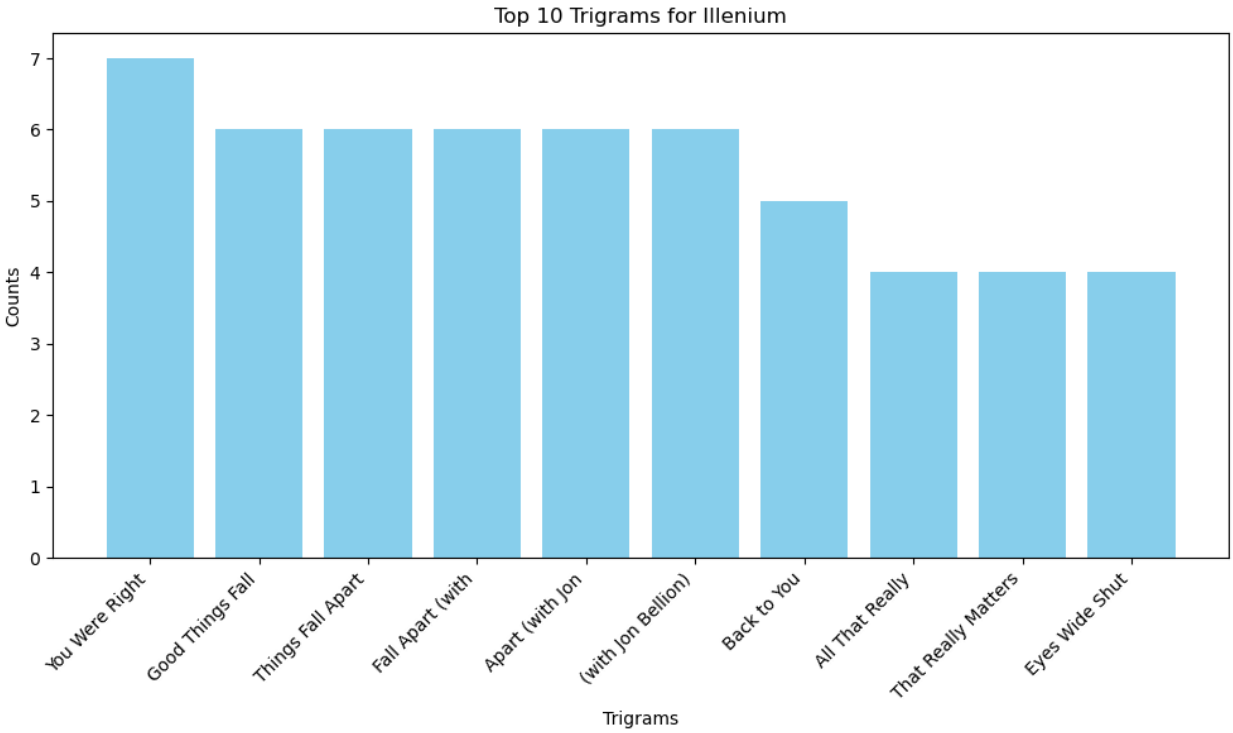
Trigrams

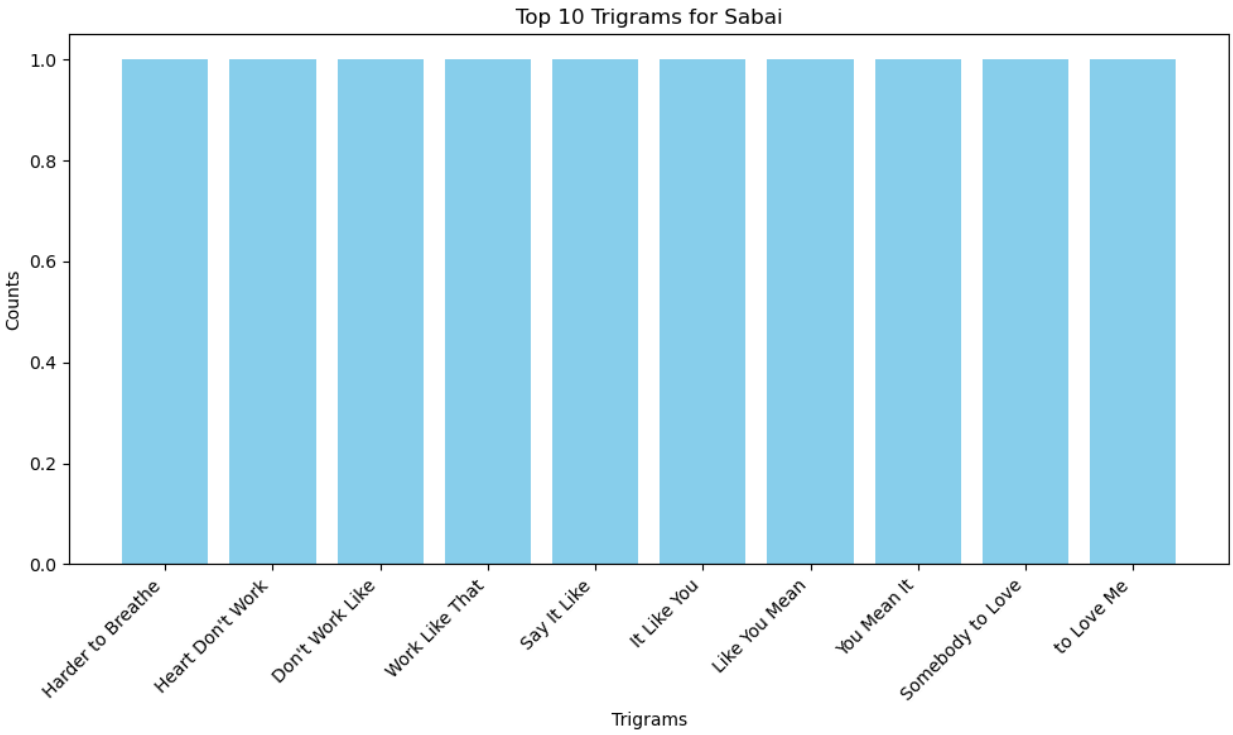
In a similar vein to how bigrams can further provide insight into song titles, going one level up into trigrams or three-word clusters can help identify song popular song titles. By adjusting the code for the bigrams, we can select for three words instead of two. An interesting emergent property is that filtering for trigrams eliminates any collaboration artists because most people's names are only two-word clusters. While exceptions could apply, this would further isolate only song titles in the output.

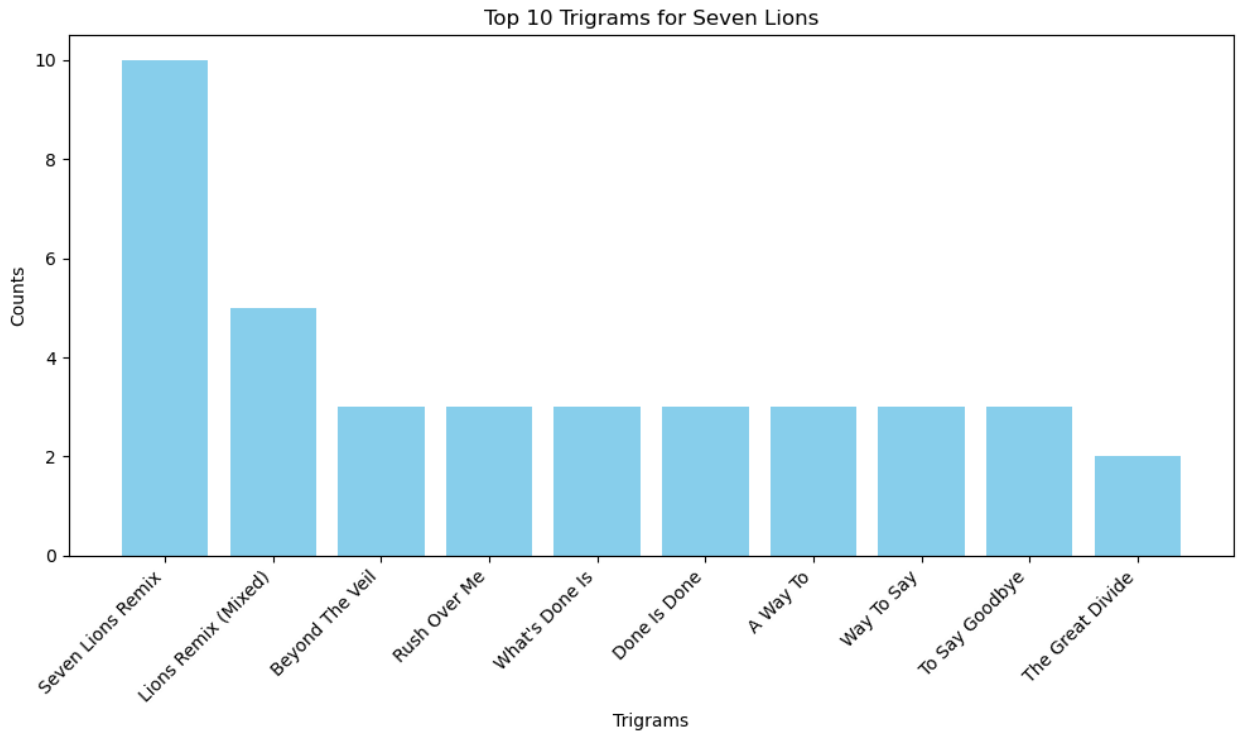
Below are our results for each artist.


Top 10 Trigrams for Dabin
Counts
Trigrams
2.00
1.75
1.50
1.25
1.00
0.75
0.50
0.25
0.00
When This Is
This Is Over
Ray Volpe Remix
Hope It Hurts
Last Heroes Remix
Skybreak VIP Remix
MYRNE VIP Remix
Rome In Silver
In Silver Remix
Digital Skies Remix


Top 10 Trigrams for Gryffin
Counts
Trigrams
2.00
1.75
1.50
1.25
1.00
0.75
0.50
0.25
0.00
Your Love (with
All You Need
If I Left
DANCE THROUGH THE
THROUGH THE NIGHT
THE NIGHT (with
NIGHT (with Norma
(with Norma Jean
Norma Jean Martine)
FALLING TO PIECES


Top 10 Trigrams for Illenium
Counts
Trigrams
0
1
2
3
4
5
6
7
You Were Right
Good Things Fall
Things Fall Apart
Fall Apart (with
Apart (with Jon
(with Jon Bellion)
Back to You
All That Really
That Really Matters
Eyes Wide Shut


Top 10 Trigrams for Sabai
Counts
Trigrams
0.0
0.2
0.4
0.6
0.8
1.0
Harder to Breathe
Heart Don't Work
Don't Work Like
Work Like That
Say It Like
It Like You
Like You Mean
You Mean It
Somebody to Love
to Love Me



Summary of Trigrams

	Result Findings
Dabin	Popular song 'When This Is Over' Ray Volpe Remix x2 Last Heroes Remix x2
Gryffin	Popular song "All You Need to Know" Popular song "If I Left The World"
Illenium	Popular song "Good Things Fall Apart with Jon Bellion"
Sabai	N/A
Seven Lions	Seven Lions Remix Popular song "What's Done is Done" Popular song "A Way to Say Goodbye"

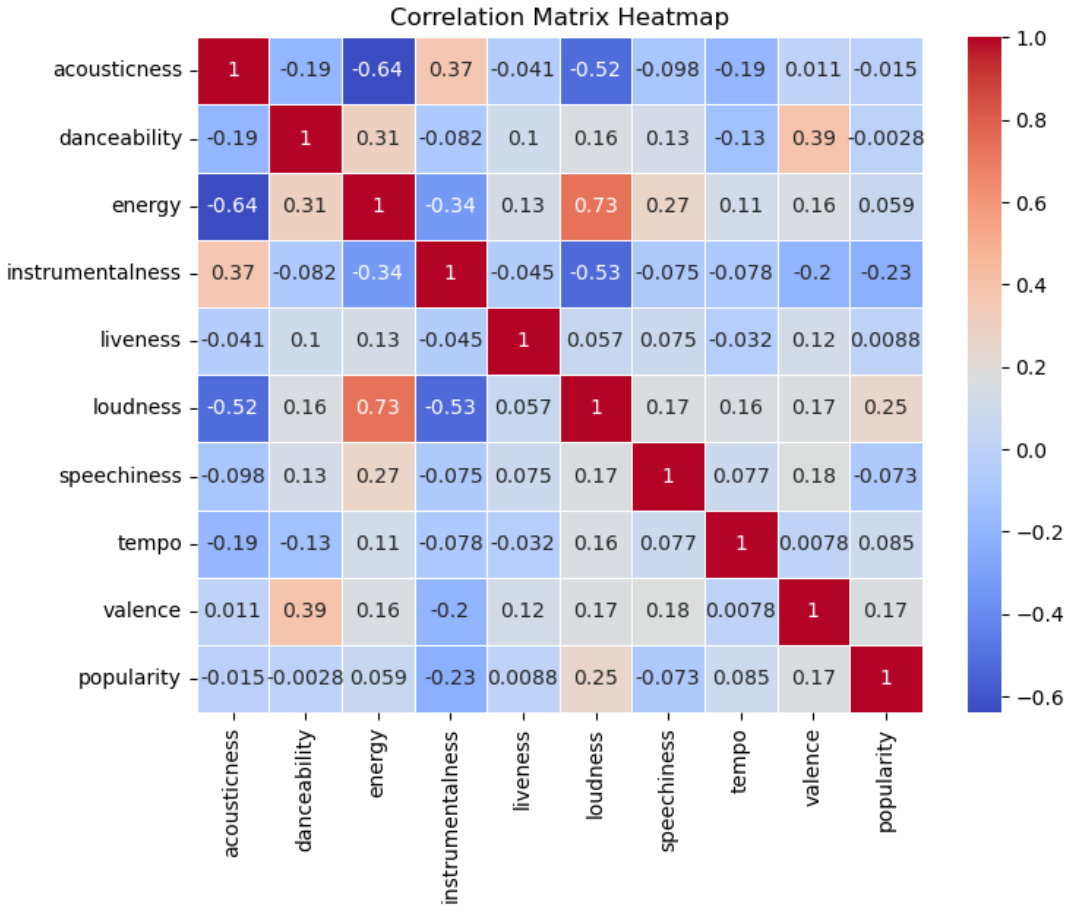
It should be noted that trigrams do not completely visualize the entirety of the song title. For instance, Illenium's 'Good Things Fall Apart with Jon Bellion' appear as consecutive level bars because it is all part of the same song. While our findings for trigram are not much extensive compared to unigram or bigram, it is easier to find out which songs have multiple versions. Other interesting tidbits extracted involve further specifics like "Last Heroes Remix" or "Ray Volpe Remix."

Regression Analysis of Sound Characteristics

In this part of the study, we look at how the characteristics of the sound might affect the song popularity. The sound data was collected from spotify and aggregated into a single data frame. In this section, we are attempting to review and analyze the sound features to see if any have a significant impact on the popularity of the song.

Correlation

First, we will investigate if any features are correlated. This may be done through a correlation matrix.



The correlation matrix shows how often each of the factors appear side by side. While it does not tell us much about this data at this point, it is good to see whether or not we need to investigate for any potential multicollinearity.

The heatmap is a great way to identify any feature-pairs that might present a problem. Positively correlated pairs lean towards dark red, while negatively correlated pairs lean towards dark blue. The deeper the color, the stronger the magnitude. A perfect correlation has a value of 1.0, which means it is essentially the same. We can easily see this demonstrated in our chart. An equivalence line in dark red cuts the chart in half diagonally and reflects how we are mapping the variables onto itself for comparison. Exploring our features in our correlation chart, we can see that loudness and energy are very closely correlated (0.73). Energy and acousticness are also somewhat strongly negatively correlated (-0.64). This suggests that we should explore potential multicollinearity before modeling any data.

Variance Inflation (VIF)

To do so, we have to check VIF score. VIF (Variance Inflation) is a measure that "gauges how much a feature's inclusion contributes to the overall variance of the coefficients of features" (Nair, A. (2022, January 4). In other words, it measures how much the variable inflates the model and how unstable it makes it. A VIF score of 1 means that feature is completely independent and has no correlation to any other features. A VIF between 1 and 5 is an acceptable range. A VIF between 5 and 10 suggests high correlation. And finally, VIF above 10 is problematic and suggests multicollinearity.

When we first run the VIF test with all the predictors, we got the following result.

	Independent Variable	Variance Inflation Score
0	acousticness	2.168464
1	danceability	21.851194
2	energy	29.069302
3	instrumentalness	1.626919
4	liveness	2.947130
5	loudness	9.880011
6	speechiness	2.588435
7	tempo	19.464955
8	valence	5.390025

As we can see, there are several features that have high VIF score. Energy, danceability, and tempo are all high. While the VIF does show how much inflation these variables provide, it does not say which other variables are related. The best method to

tackle this is to remove the variables one by one and re-run the VIF test, starting with the highest VIF score. In this case, we will remove 'energy' because it has the greatest VIF score.

	Independent Variable	Variance Inflation Score
0	acousticness	2.056357
1	danceability	14.250346
2	instrumentalness	1.626869
3	liveness	2.848591
4	loudness	9.559906
5	speechiness	2.487399
6	tempo	11.660797
7	valence	5.388931

Next, we continue to remove the next highest feature again because there are features with high VIF score still. Following our method, we will remove 'danceability' which has a score of 14, still higher than our acceptable 10.

After running out VIF test again, we finally find a more acceptable range of values. Note that 'tempo' and 'loudness' are still over the typical safe range of 1 to 5. We will keep this in mind as we build our model.

	Independent Variable	Variance Inflation Score
0	acousticness	1.974014
1	instrumentalness	1.626849
2	liveness	2.771109
3	loudness	8.634491
4	speechiness	2.448150
5	tempo	9.305202
6	valence	4.188920

To begin building our model, we will first try lasso regression. Lasso involves use of absolute values to penalize, or regularize, the coefficients. Regression with many predictors tend to overfit the model, and so, lasso would be one good way to control for this tendency.

Lasso

We first try out lasso without normalization and with normalization. When implementing lasso, we will use only the predictors that have passed our VIF test. This is to ensure that no features are over-represented in the model through multicollinearity.

Normalization is a process where we standardize all the data to control for the metric bias inherent in different counters. There are two different ways to normalize: min-max scaling and z-score. For this model, we will choose to use the z-score, which centers all the data around its respective z-score. The table below summarizes the model with our results.

Lasso without Normalization	Lasso with Normalization
R Squared Value: 0.05 Mean absolute error: 10.89 Mean squared error: 163.43 Root mean squared error: 12.78	R Squared Value: 0.06 Mean absolute error: 0.83 Mean squared error: 0.96 Root mean squared error: 0.98

The r-square value indicates how well the model explains the variability of the outcome data. An r-squared of 1 says the model perfectly predicts the data while an r-squared of 0 means it does not explain any of the variability of the data. We thus want a higher r-squared value. From our table, we can see that Lasso without normalization I slightly worse than Lasso with Normalization because the r-squared value is lesser. While both r-squared values are really low, we are only reviewing the model comparisons at this point.

Mean absolute error, mean squared error, and root mean squared error are all different metrics that assess the same thing. They all measure the difference between the predictive value and the actual value of the outcome variable (which in this case is the popularity of the songs). Because of this, by heuristics, smaller values imply better fit. We want an ideal (though hardly ever true) value of 0, which would say that there is no difference between the model and reality.

Looking at the two models, Lasso without Normalization is clearly higher on all values of error. 10.89 for MAE, 163 for MSE, and 12 for RMSE. Once normalized, these errors shrink to 0.83, 0.96, 0.98 respectively. It is thus clearly obvious that Lasso without Normalization is a superior model on all counts.

Recall that we preserved two potentially high VIF factors: Tempo and Loudness. To see if we can get an even better model, we will attempt to remove each sequentially (by VIF value) and run the Lasso model again. In line with our earlier finding, we will use only the normalized dataset. The results are as follows.

Test: Assumption of additional VIF influence (normalized)

Removal of Tempo

	Independent Variable	Variance Inflation Score
0	acousticness	1.855998
1	instrumentalness	1.602997
2	liveness	2.649879
3	loudness	5.043427
4	speechiness	2.264694
5	valence	3.603767

```
R Squared Value: 0.05
Mean absolute error: 0.84
Mean squared error: 0.96
Root mean squared error: 0.98
```

Removal of Loudness

	Independent Variable	Variance Inflation Score
0	acousticness	1.558349
1	instrumentalness	1.320729
2	liveness	2.355592
3	speechiness	2.189727
4	valence	3.003100

```
R Squared Value: 0.01
Mean absolute error: 0.86
Mean squared error: 1.01
Root mean squared error: 1.00
```

When tempo was removed, our r-squared value went down while the error metrics stayed relatively the same. This means the model without tempo may be comparable to our previous model. Removing the next highest VIF feature loudness, we see that our r-squared took a rapid dive from 0.05 to 0.01. This is a strong sign that the model has gotten worse. The error metrics also properly reflects this as well with higher error rates.

For curiosity, we also ran the test with all multicollinear features in spite of VIF factors. The result model statistics are below.

Test: Lasso with all predictors (normalized)

```
R Squared Value: 0.03
Mean absolute error: 0.84
Mean squared error: 0.99
Root mean squared error: 0.99
```

As expected, a model with all predictive features are also not good. The r-squared value is below optimum compared to our previous models and the error rates are a lot higher, especially the MSE and RMSE.

Considering all the models above, we will have the following conclusion. The best model for Lasso is with all the predictors without energy and danceability.

Conclusion for Lasso
Best model is with the following predictors: 'acousticness','instrumentalness','liveness', 'loudness', 'speechiness', 'tempo', 'valence'

Ridge

Next, we will explore the regression analysis with ridge. Like lasso, ridge controls for predictors and involves a penalty to limit the impact from overfitting. However, unlike lasso, ridge involves use of squares and minimizes factors to reduce impact rather than remove it. We use the basic framework of lasso to create the ridge regression and find the following.

Ridge		
All Predictors	Selected Predictors	Selected Predictors
R Squared Value: 0.01 Mean absolute error: 0.84 Mean squared error: 1.00 Root mean squared error: 1.00	R Squared Value: 0.05 Mean absolute error: 10.90 Mean squared error: 164.08 Root mean squared error: 12.8	R Squared Value: 0.05 Mean absolute error: 0.84 Mean squared error: 0.96 Root mean squared error: 0.98
Normalized	Not normalized	Normalized

For control, we will look at the normalized model with our previously selected predictors.

Comparing the selected model to a normalized model with all predictors, we can see that our selected model outperforms the other by a low. The r-squared value is 5 times greater than the model without selected features and the error rate is lower for MSE and RMSE.

With contrast to a not normalized model of the selected model, the normalized model also performs better, especially in having smaller error rates.

Selected Lasso	Selected Ridge

R Squared Value: 0.06 Mean absolute error: 0.83 Mean squared error: 0.96 Root mean squared error: 0.98	R Squared Value: 0.05 Mean absolute error: 0.84 Mean squared error: 0.96 Root mean squared error: 0.98

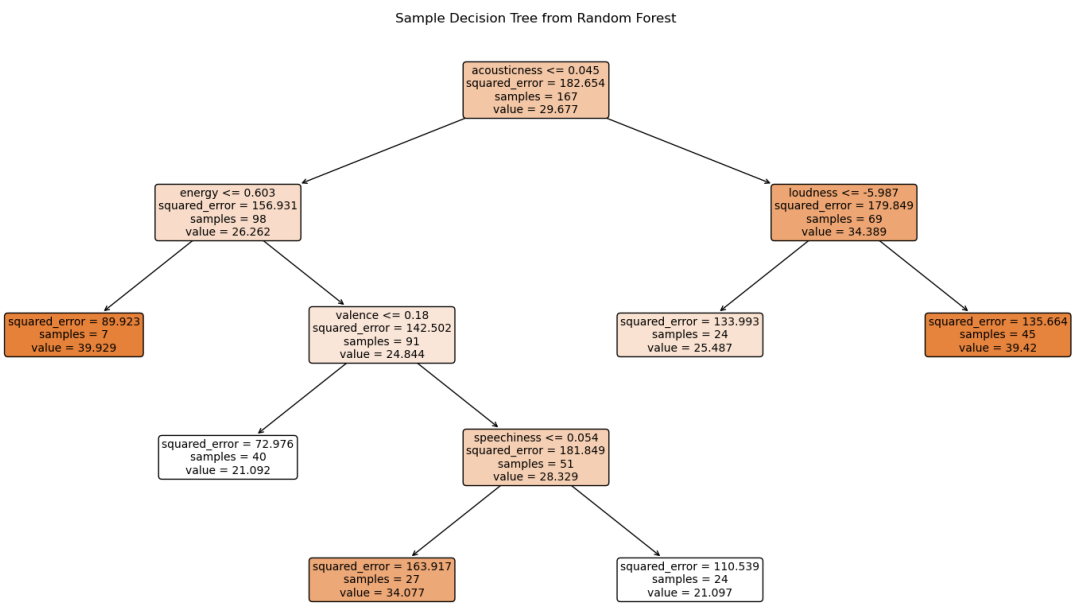
Comparing our two regressions, both models are comparable. However, lasso might be a better model for this data as it has a higher r-squared value and less MAE. This may be because lasso removes the unnecessary predictors entirely while Ridge retains it (in a very minimal amount).

Random Forest

Random forest is an ensemble machine learning algorithm that combines predictions from multiple models of decision trees to improve accuracy. Decision trees involve splitting data by feature values and then creating classification based on the split. These "decisions" are also optimized to produce the most accurate results. Multiple trees are used in an ensemble to produce a robust model like random forest.

Because random forest does not rely on numbers or calculations of values like the typical regression (like ordinary least squares), it does not need to be normalized. It also allows the use of all predictors because VIF is not a factor that affects random forest because it relies on splitting thresholds and not computing linear relationships.

Below is an example tree taken from random forest to illustrate how random forest works.



Running this ensemble involves a few steps.

The first step is to split our data set into training and testing. We will use a conventional 60:40 split, where 60% of the data is used for training and 40% is used to test (or choose) the model. Because we plan to use GridSearch, we do not need to use a validation set for hyperparameter tuning.

The next step, we will use sklearn's library for GridSearchCV() and RandomForestRegressor(). The purpose of grid search is hyperparameter tuning. The regressor relies on a few parameters (that are external to this models' data) to base its decisions and analysis. For this reason, we use grid search to find the optimum values for these hyperparameter. Once we find the optimum parameters to tune the regressor, we run the regressor algorithm RandomForestRegressor(). A side note is that we are using RandomForestRegressor() because the outcome variable is a numeric and has a range of values, whereas we would use RandomForestClassifier() if it was a binary value of 1 or 0.

Once we run the algorithm, we find the results for this model's fit below,

Random Forest Regression
Mean Absolute Error: 9.75 Mean Squared Error: 142.75 Root Mean Squared Error: 11.95 R^2: 0.11

This is surprising. Contrast to our previous Lasso model, random forest has better results. A side-by-side comparison shows that the r-squared value goes up from 0.06 to 0.11. This is a significant improvement. The error scores for this value cannot be properly compared because the random forest used a non-normalized dataset, but regardless, it still shows good promise of better model performance. Below is a side by side between Lasso (not normalized for comparison) and Random Forest.

Lasso (not Normalized)	Random Forest Regression
R Squared Value: 0.05 Mean absolute error: 10.89 Mean squared error: 163.43 Root mean squared error: 12.78	Mean Absolute Error: 9.75 Mean Squared Error: 142.75 Root Mean Squared Error: 11.95 R^2: 0.11

As such, we will base our model on the ensemble method of random forest because it provides the optimal value in assessing the data. A deeper look at the ensemble features reveal the following, ordered by most importance. This can accessed with the feature_importance_.

	Feature	Importance
5	loudness	0.337883
0	acousticness	0.284660
8	valence	0.098985
2	energy	0.087788
3	instrumentalness	0.057736
6	speechiness	0.038409
4	liveness	0.034858
1	danceability	0.033822
7	tempo	0.025858

Thus for this music subgenre of melodic electronic, loudness and acousticness seems to be a prevalent feature tied to popularity. Artists still developing their foothold like Sabai may opt to make their music emulate these features to capture more fans.

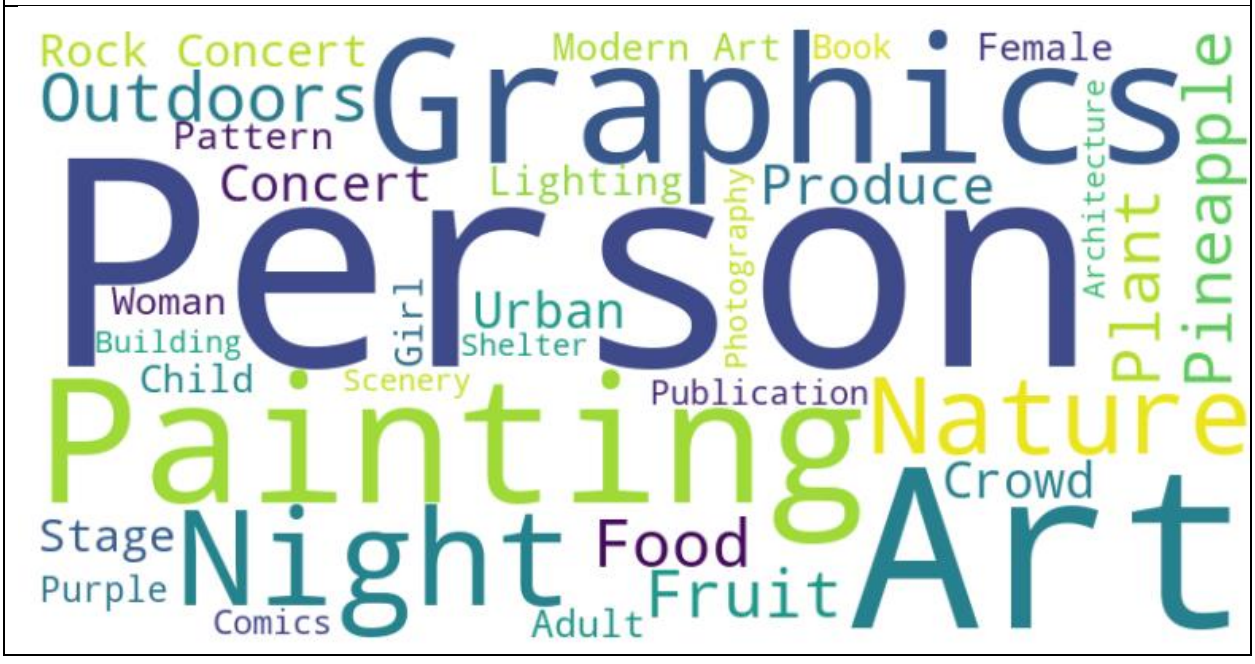
General Overview of Cover Art

Traditionally cover art were the selling point of the artists. Similar to how book covers provide a draw to prospective consumers, cover art plays a role in mass appeal. In the contemporary music landscape, cover art may serve as a representation of music, especially for tracks without any music videos or traditional visuals. In this section of the analysis, we use AWS Rekognition to analyze the images and provide labels with confidence scores of the selected artists. For this analysis, we only select labels that have a confidence score about 90%. We then combine this data set with Spotify's scraped data to review how cover art (and its characteristics) influence popularity.

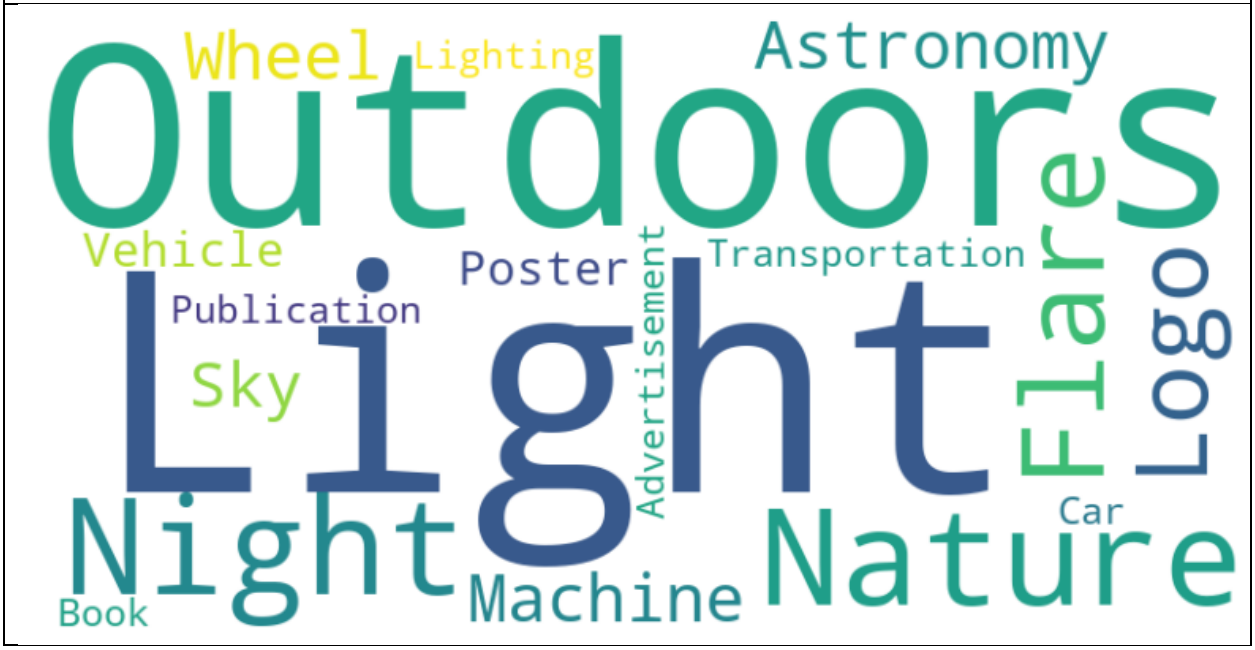
Cover Art Trends

We start with a general descriptive capture of each artist and the associated features related to their cover art. To do this, each artist is filtered and their associated Rekognition labels are combined into a dictionary and outputted as a word cloud. Below are the following artists and their associated word cloud.

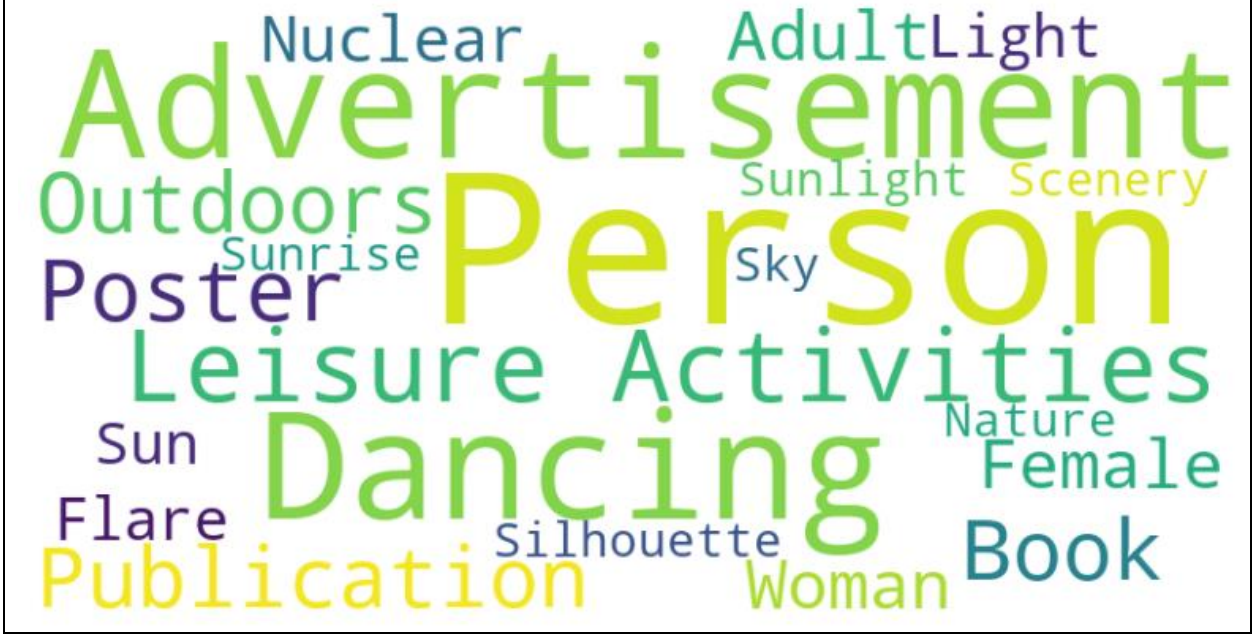
Dabin



Gryffin

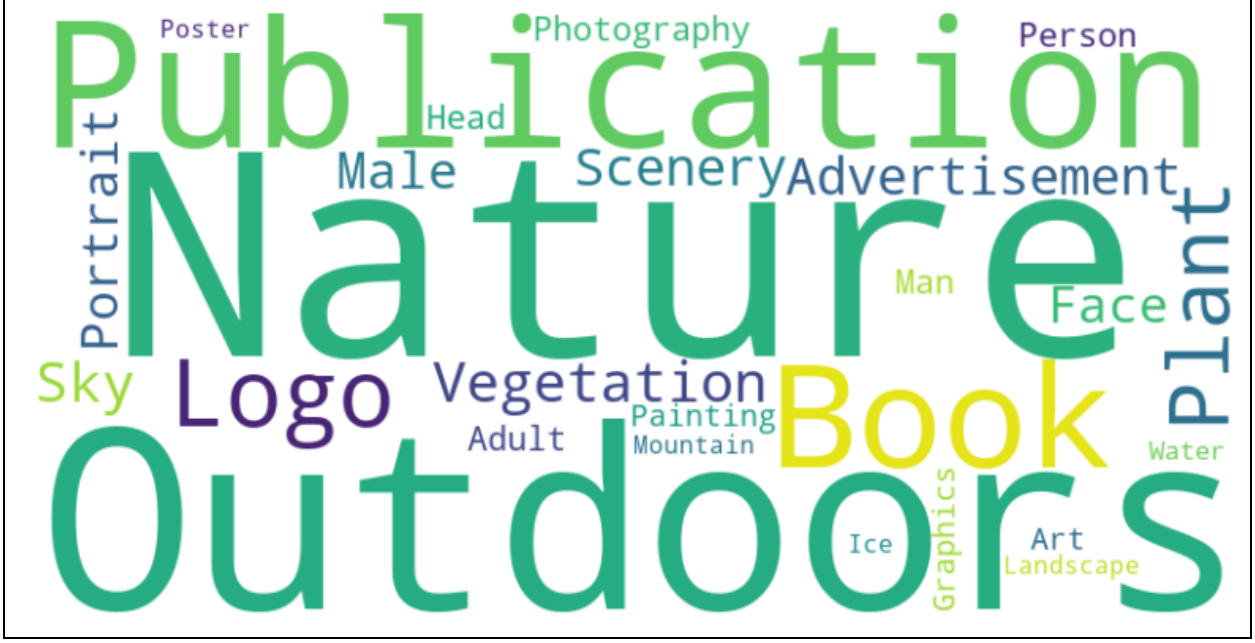


Illenium



Sabai

Seven Lions



A summary of each artist can be see below. While the word cloud offers a more comprehensive look, I will only add a few of the top re-occurring labels as provided by Rekognition. Highlighted labels appear more than once across the table.

Comparative Similarities among Artists

Dabin	Gryffin	Illenium	Sabai	Seven Lions
Person	Light	Person	Purple	Nature
Art	Outdoors	Advertisement	Plant	Outdoors
Painting	Night	Dancing	Vegetation	Publication
Graphics	Nature	Leisure		Book
Night	Flare	Activities		Logo
Nature	Logo	Poster		Plant
Outdoors	Astronomy	Book		Vegetation
Fruit	Wheel	Publication		Sky
Pineapple	Machine	Outdoors		Advertisement
Food	Sky	Adult		Male
Plant	Poster			Face
Produce				
Crowd				

As you can see from the comparative table above, many of these themes intersect within this music subgenre. Outdoors and Nature seems to be a common theme. This resonates with the music crowd audience, who likes to enjoy outdoor concerts and festivals. Many of the cover art is also very stylistic or professional, which might provide insight as to why Poster, Publication, Book, and Advertisement might be a common trait that appears as well.

An interesting note is to see that Sabai has the least number of labels. This could be reflective of his comparative newness to the industry, and hence having less albums and less descriptive labels.

A deeper in-depth look at the commonalities can be seen below. This information can be extracted by filtering the artists, combining values into a dictionary, and then printing out the associated key:value pairs per length.

```
Labels with four artists

Nature: ['dabin', 'seven_lions', 'gryffin', 'illenium']
Outdoors: ['dabin', 'seven_lions', 'gryffin', 'illenium']
Book: ['dabin', 'seven_lions', 'gryffin', 'illenium']
```

Publication: ['dabin', 'seven_lions', 'gryffin', 'illenium']

Labels with three artists

Person: ['dabin', 'seven_lions', 'illenium']
Plant: ['dabin', 'seven_lions', 'sabai']
Adult: ['dabin', 'seven_lions', 'illenium']
Scenery: ['dabin', 'seven_lions', 'illenium']
Advertisement: ['seven_lions', 'gryffin', 'illenium']
Poster: ['seven_lions', 'gryffin', 'illenium']
Sky: ['seven_lions', 'gryffin', 'illenium']

Labels with two artists

Art: ['dabin', 'seven_lions']
Painting: ['dabin', 'seven_lions']
Graphics: ['dabin', 'seven_lions']
Night: ['dabin', 'gryffin']
Lighting: ['dabin', 'gryffin']
Woman: ['dabin', 'illenium']
Female: ['dabin', 'illenium']
Purple: ['dabin', 'sabai']
Photography: ['dabin', 'seven_lions']
Flare: ['gryffin', 'illenium']
Light: ['gryffin', 'illenium']
Vegetation: ['seven_lions', 'sabai']
Logo: ['seven_lions', 'gryffin']

Unique labels

Fruit: ['dabin']
Pineapple: ['dabin']
Food: ['dabin']
Produce: ['dabin']
Crowd: ['dabin']
Concert: ['dabin']
Urban: ['dabin']
Rock Concert: ['dabin']
Stage: ['dabin']
Girl: ['dabin']
Pattern: ['dabin']
Modern Art: ['dabin']
Child: ['dabin']
Comics: ['dabin']
Shelter: ['dabin']
Building: ['dabin']
Architecture: ['dabin']
Dancing: ['illenium']
Leisure Activities: ['illenium']
Nuclear: ['illenium']
Sun: ['illenium']
Sunlight: ['illenium']
Silhouette: ['illenium']

```
Sunrise: ['illenium']
Male: ['seven_lions']
Face: ['seven_lions']
Portrait: ['seven_lions']
Head: ['seven_lions']
Man: ['seven_lions']
Water: ['seven_lions']
Landscape: ['seven_lions']
Ice: ['seven_lions']
Mountain: ['seven_lions']
Astronomy: ['gryffin']
Wheel: ['gryffin']
Machine: ['gryffin']
Vehicle: ['gryffin']
Transportation: ['gryffin']
Car: ['gryffin']
```

A review of the top common features seems to be nature, sky, photography, and design. Some of the more differentiating features seem to be more concrete things like wheels, machines, architecture, fruit, pineapple, ice, etc. A summary is tabled below.

Comparative Differences among Artists

Dabin	Gryffin	Illenium	Sabai	Seven Lions
Fruit	Astronomy	Nuclear	N/A	Male
Pineapple	Wheel	Sun		Face
Food	Machine	Sunlight		Portrait
Produce	Vehicle	Silhouette		Head
Crowd	Transportation	Sunrise		Man
Concert	Car			Water
Rock band				Landscape
Girl				Ice
Pattern				Mountain
Modern Art				

Sabai has no unique themes, likely due to his inherent inexperience and likeliness to follow the industry norms. More experienced artists like Seven Lions or Dabin are likely to venture into new cover styles to attempt to differentiate and cast a wider audience appeal.

Designs by Popularity

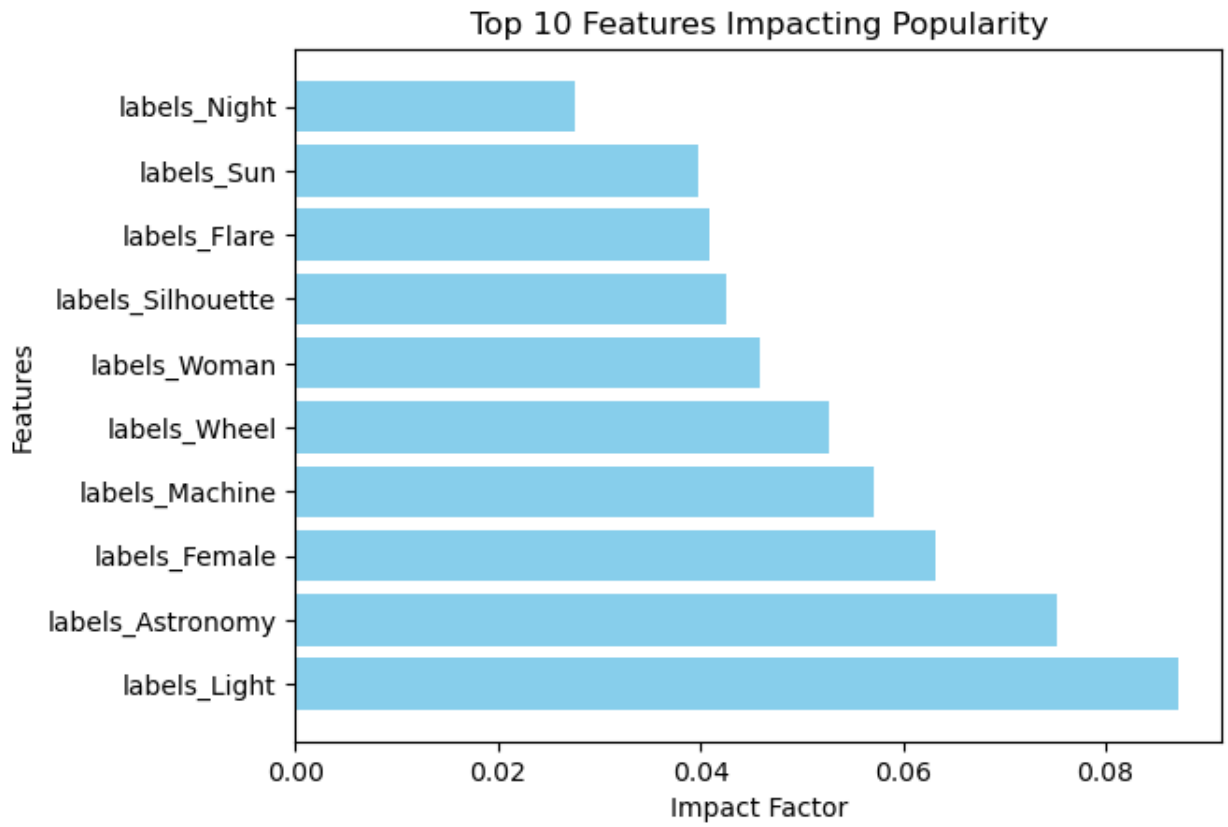
We previously compare and contrast trends between the artists, but we ultimately want to know how cover art can translate into popularity and capitalize on it. While songs are not entirely dependent on cover art and many factors that directly affect song popularity are excluded from this research, the goal of this analysis is to see how each cover art or style choice may influence a song's popularity.

Similar to our earlier ensemble method, we will use random forest again to investigate the relationship between the cover art and the popularity associate with each song. First, we use the comprehensive data set that combined the Rekognition labels with the information scraped from Spotify. We then split this database into a 60:40 training-test. We then find the hyperparameter with GridSearchCV and then use these inputs to develop our ensemble method RandomForestRegressor(). We then train the regressor on the training set and run the trained regression gain with test set. Our output is as follows.

Performance Metrics for Cover Art

```
[62 rows x 2 columns]
Mean Absolute Error: 10.16
Mean Squared Error: 146.57
Root Mean Squared Error: 12.11
R²: 0.18
```

While this model does not do a great job at capturing 82% of the variance, we can still glean off information from this model – namely the feature importance. By pulling the feature importance from the ensemble and sorting, we can extract the top ten features that affects the popularity. Below is the summary chart.



From this diagram, we can see that the top feature in the horizontal bar graph is Light, followed by astronomy and female. This may tell us more about the audience base. Concerts for melodic electronic music is typically held outdoors at night, with an emphasis on lights. This may be indicative or reflective of how the audience consumes the music. The interesting feature that appears here is Female, which may suggest that female might be a large makeup of this audience base.

As such, newer artists looking to break into the melodic electronic music subgenre should focus on concert aesthetics targeting female listeners, especially when designing their cover art.