Spotify's Most Streamed Songs in 2023

Data Science Team:
Gavin Grosswald
Erika Haase
Maybel Herrera
Anthonyo Kinch-Rice

Syracuse University
IST 687: Introduction to Data Science
Professor Toni Hanrahan

Abstract

The music industry and music production process are constantly evolving, driven by the widespread adoption of music streaming platforms and the increase in social media marketing. Music now has the capability to quickly reach all audiences and different demographics across the entire world. The growth in technology has allowed for faster output of music, seamless editing, and widespread traction.

Table of Contents

.

1.1 Introduction

The intent of this project is to understand how to interoperate and manipulate large data sets, specifically datasets with 10,000 data points or more.

The Project Team has assimilated into the role of data analysts working for Produce that Beat ("the Label"), an innovative record label poised to revolutionize the music industry. Produce that Beat has the tenacity and drive to be a disrupter in the industry, and the Label has recruited the data analysis team to help develop a new strategy to position the Label for immediate success. Historically, the Label has relied on its Agents' experience in the industry, and understanding of the characteristics that lead to success. While this has been fruitful for the Label, it has certainly come with its challenges and occasional failures. At Produce that Beat's 2023 annual Board of Directors meeting, a new strategy was developed to help the Label compete while minimizing additional risk.

The Label provided the data analysis team with a Spotify dataset to help guide the decision-making process. By strategically leveraging Spotify data of the top 1,000 artists and songs, the Label aims to identify emerging artists and produce chart-topping songs. By combining data analytics with a deep understanding of music trends, the Label believes it can establish itself as a leader in the industry, paving way to a new era of talent discovery and music production.

Produce that Beat has a two-fold data strategy: identify emerging artists and produce chart-topping songs.

Produce That Beat aims to streamline the process of discovering the next big artists. By carefully analyzing Spotify data, including top charts, streaming patterns, and listener preferences, the Label will identify promising musicians poised for success. This data-driven approach will allow the Label to make informed decisions in talent acquisition, ensuring the hiring of artists with the potential for success.

Produce That Beat will leverage Spotify data to guide the music production process. By identifying patterns in successful songs, analyzing key elements that resonate with audiences, and staying attuned to evolving musical preferences, the Label will aim to produce songs that have an increased likelihood of topping the charts. This strategy positions the Label to consistently deliver content that aligns with current market demands while maximizing the Label's success in the competitive music landscape.

2.1 The Data Set

The Label leveraged a dataset from kaggle.com, Most Streamed Spotify Songs 2023, which includes 1,000 of the top streaming songs from Spotify and a comparison of those Spotify songs to other competitive streaming platforms like Apple and Deezer. The dataset is a comprehensive list of the most popular songs of 2023 as determined by the Spotify music streaming platform. The dataset includes key metrics (listed below) that apply to each of the 1,000 songs which the Label was able to analyze to drive its business strategy forward. The data set contains 24 variables, 953 observations, and 22,872 total data points. Key attributes of the data set include:

- Song Title: Name of the song
- Artist Name: Name of the artist(s) of the song
- Artist Count: Number of artists contributing to the song
- Song Release Year: Year when the song was released
- Song Release Month: Month when the song was released
- Song Release Day: Day of the month when the song was released
- Spotify Playlist Count: Number of Spotify playlists the song is included in
- Spotify Charts Count: Presence and rank of the song on Spotify charts
- Total Streams: Total number of streams on Spotify
- Apple Playlist Count: Number of Apple Music playlists the song is included in
- Apple Charts Count: Presence and rank of the song on Apple Music charts
- Deezer Playlist Count: Number of Deezer playlists the song is included in
- Deezer Charts Count: Presence and rank of the song on Deezer charts
- Shazam Charts Count: Presence and rank of the song on Shazam charts
- BPM: Beats per minute, a measure of song tempo
- Key: Key of the song
- Mode: Mode of the song (major or minor)
- Danceability: Percentage indicating how suitable the song is for dancing
- Valence: Positivity of the song's musical content
- Energy: Perceived energy level of the song
- Acousticness: Amount of acoustic sound in the song
- Instrumentalness: Amount of instrumental content in the song
- Liveness: Presence of live performance elements
- Speechiness: Number of spoken words in the song

3.1 Cleaning the Data

Cleaning a dataset can be one of the more complex tasks in any data science project. As part of the analysis, Produce the Beat found that much of the sample size was usable. The team employed data munging techniques to clean the raw data and transform it into multiple different data frames to serve as the foundation for all of Produce that Beats analysis.

Column adjustments included renaming all variables to be more descriptive, removing columns that were not considered applicable to the analysis, and reordering the column structure to create a more appropriate flow of information.

This was followed by combining the release day, month, and year columns into one release date column, and creating a new column that calculates how many days it has been from today since the song was released. The team also converted character variables into numeric variables to allow for the use of mathematical calculations.

There were several issues with song names including unusable characters that had to later be excluded for text mining analysis and the review of the words that make up the most popular songs. These unusable characters however, did not impact any quantitative analysis.

4.1 Data Analysis

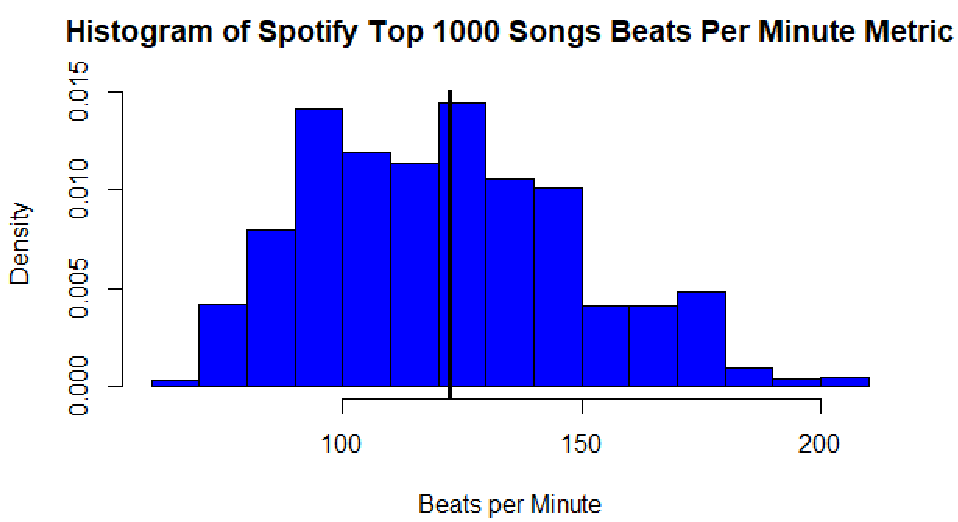
The goal of the Label's analysis is to understand the components and attributes that lead to success from both the talent discovery and music production perspective. How can Produce that Beat leverage its data insights to position itself as a leader within the industry while continuously driving results for its artists and new music?

After reviewing the entire raw dataset, the team decided to focus its analysis on just a few variables: Beats per Minute, Energy, Speechiness, and Acousticness ("focus group variables").

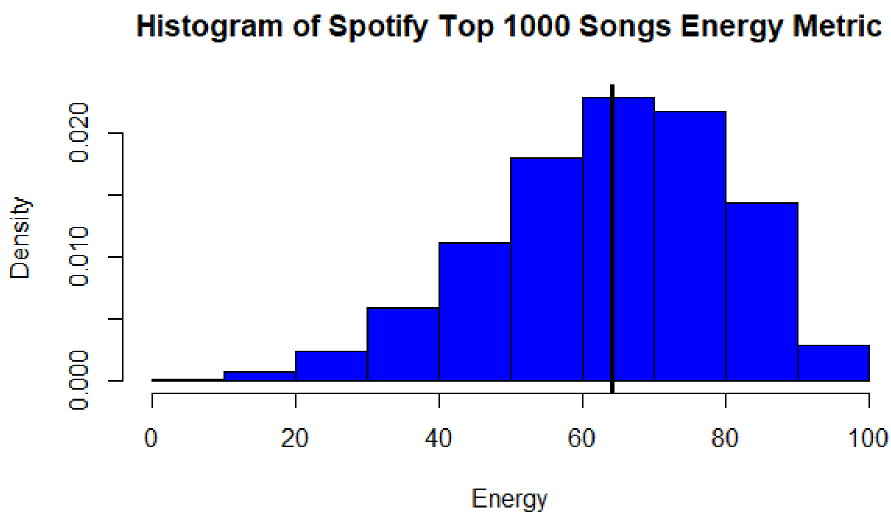
Below, a detailed analysis of the Spotify dataset:

4.2 Histograms of Key Metrics

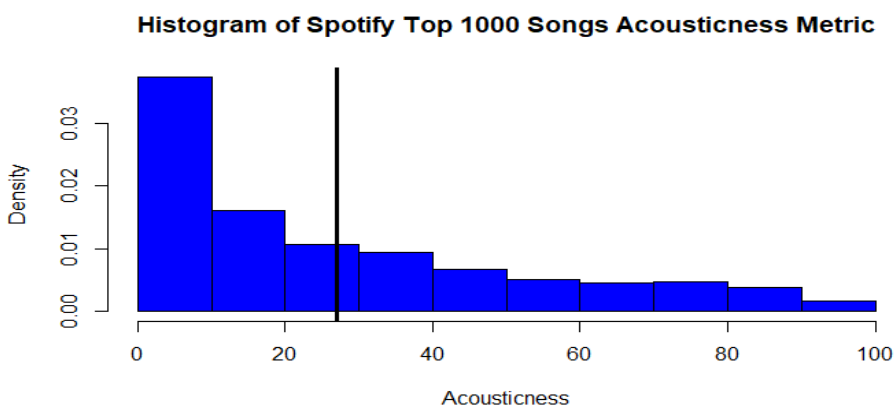
Beats per Minute measures the tempo, or speed, of a song and the histogram below indicates that on average, songs have a BPM of 122.5 with peaks around 100 and 125. BPM is normally distributed around the mean and has a range of 206 BPM less 65 BPM.



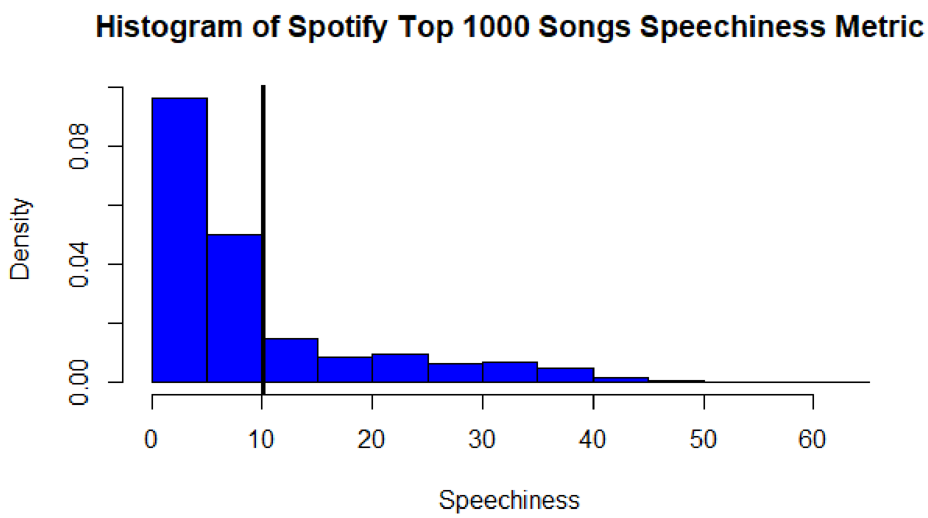
Energy, which refers to the perceived energy level of a song, is similarly normally distributed around the mean of 64%. Energy had quite a large range of 9% to 98% but the chart does have a slight left skew tendency meaning most songs are on the higher Energy level.



Acousticness is the amount of acoustic within a song. This histogram is interesting as it shows a consistent decline with an average of 27%. This chart indicates that while acoustic music can still be popular, consumer preferences are clearly on the lower acoustic spectrum.



Speechiness is the number of spoken words within a song. This histogram indicates that 50% of songs have Speechiness levels of 10% or less. The median is 6%. This shows that listeners prefer fewer and more repetitive words. Examples of low Speechiness include Coldplay and Taylor Swift with simple repeated choruses. However, there are hits with high Speechiness. Most high Speechiness songs include rap lyrics with high wordcounts by artists like Drake and Nicki Minaj.



4.3 Correlation Analysis

An initial correlation analysis was ran using all song metrics that impact total streams and results can be seen in the below table:

	Total Streams	Beats per Minute	Danceability	Valence	Energy	Acousticness	Instrumentalness	Liveness	Speechiness
Total Streams	1.00	0.00	-0.10	-0.04	-0.03	0.00	-0.04	-0.05	-0.11
Beats per Minute	0.00	1.00	-0.15	0.04	0.03	-0.02	0.00	0.00	0.04
Danceability	-0.10	-0.15	1.00	0.41	0.20	-0.24	-0.09	-0.08	0.18
Valence	-0.04	0.04	0.41	1.00	0.36	-0.08	-0.13	0.02	0.04
Energy	-0.03	0.03	0.20	0.36	1.00	-0.58	-0.04	0.12	0.00
Acousticness	0.00	-0.02	-0.24	-0.08	-0.58	1.00	0.04	-0.05	-0.02
Instrumentalness	-0.04	0.00	-0.09	-0.13	-0.04	0.04	1.00	-0.05	-0.08
Liveness	-0.05	0.00	-0.08	0.02	0.12	-0.05	-0.05	1.00	-0.02
Speechiness	-0.11	0.04	0.18	0.04	0.00	-0.02	-0.08	-0.02	1.00

Stronger correlations found between Energy and Acousticness (-0.58), Danceability and Valence (0.41), Valence and Energy (0.36), and Danceability and Acousticness (-0.24). The negative correlation between Energy and Acousticness is most interesting as both variables were in our focus group. The correlation indicates that as the percentage of Energy in a song increases, the percentage of Acousticness decreases.

After reviewing the raw correlation analysis, the team decided to run it again including only the focus variables, Total streams, BPM, Energy, Acousticness, and Speechiness. As expected, the results were the same as the quantity of variables does not impact individual correlations.

	Total Streams	Beats per Minute	Energy	Acousticness	Speechiness
Total Streams	1.00	0.00	-0.03	0.00	-0.11
Beats per Minute	0.00	1.00	0.03	-0.02	0.04
Energy	-0.03	0.03	1.00	-0.58	0.00
Acousticness	0.00	-0.02	-0.58	1.00	-0.02
Speechiness	-0.11	0.04	0.00	-0.02	1.00

Total Streams: Little correlation with the other focus variables but Speechiness has the strongest negative impact on Total Streams (-0.11). As Speechiness decreases, Total streams increases.
Beats per minute: Not correlated with any other of the focus variables meaning a songs BPM has little impact on the percentages of other metrics
Energy: Has a strong negative correlation with Acousticness (-0.58) meaning as a song's percentage of Energy increases, the Acousticness levels of the song decreases
Acousticness: As stated directly above, there exists a strong negative correlation (-0.58) with Energy
Speechiness: As stated above, there exists a slight negative correlation with Total Streams (-0.11)

4.4 Linear Regression

A multi-variable linear regression model was constructed to help understand the impact song attributes have on Total Streams. The team ran many variations of the regression model in hopes of finding the best fit equation to predict future streams. For each version, one thing was held constant, Total Streams as the dependent variable (y-variable). What changed were the independent variables (x-variable or input) and the team analyzed the changes in input on Total Streams. A 5% significance level was used.

Linear Regression Model 1: Total Streams as a result of BPM, Danceability, Valence, Energy, Acousticness, Instrumentalness, Liveness, and Speechiness

$$\textbf{Total Streams} = \beta 0 + \beta 1\textbf{BPM} + \beta 2\textbf{Danceability} + \beta 3\textbf{Valence} + \beta 4\textbf{Energy} + \beta 5\textbf{Acousticness} + \beta 6\textbf{Instrumentalness} + \beta 7\textbf{Liveness} + \beta 8\textbf{Speechiness}$$

```
Coefficients:
                   Estimate Std. Error t value Pr(>|t|)    
(Intercept)      1025766945  168939782   6.072 1.83e-09 ***
Beats_per_Minute    -287588     662372  -0.434  0.66426    
Danceability       -4129475    1462378  -2.824  0.00485 ** 
Valence              157549     930743   0.169  0.86562    
Energy             -1078574    1470155  -0.734  0.46335    
Acousticness       -1076596     894483  -1.204  0.22905    
Instrumentalness   -4282274    2193162  -1.953  0.05117 .  
Liveness           -2504702    1344580  -1.863  0.06280 .  
Speechiness        -5688567    1878853  -3.028  0.00253 ** 
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 5.61e+08 on 944 degrees of freedom
Multiple R-squared:  0.02877,	Adjusted R-squared:  0.02053 
F-statistic: 3.495 on 8 and 944 DF,  p-value: 0.0005477
```

The first thing the team noted was that the p-value was less than alpha (0.05) meaning the regression is statistically significant. Second the team noticed that the intercept was extremely high. This means that all else equal, meaning every input variable is assumed to be 0, songs automatically begin with 1.02 billion streams. The team then went through each coefficient noting which ones were significant and could be used in the model versus which ones would need to be ignored. Those with p-values less than 0.05 are said to be significant and can be used in the analysis. The team noticed that each significant input variable had a negative coefficient meaning it reduced the Total Streams from the highly inflated intercept. Lastly the team noted an adjusted R^2 value of 2.1%. This means that only 2.1% of the change in Total Streams can be attributed to the input variables used in the model. This is not very strong.

Linear Regression Model 2: Total Streams as a result of BPM, Danceability, Valence, Energy, Acousticness, Instrumentalness, Liveness, Speechiness but the intercept is set to 0

```
Coefficients:
                  Estimate Std. Error t value Pr(>|t|)    
Beats_per_Minute  2011870     553645   3.634 0.000294 ***
Danceability      1437178    1160724   1.238 0.215959    
Valence          -1249526     918379  -1.361 0.173972    
Energy            4000107    1231748   3.248 0.001205 ** 
Acousticness      1968334     754607   2.608 0.009240 ** 
Instrumentalness -3600244    2231461  -1.613 0.106991    
Liveness         -1250510    1353597  -0.924 0.355804    
Speechiness      -5794038    1914094  -3.027 0.002537 ** 
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 571500000 on 945 degrees of freedom
Multiple R-squared:  0.4461,    Adjusted R-squared:  0.4414 
F-statistic: 95.14 on 8 and 945 DF,  p-value: < 2.2e-16
```

In this second regression model, the same input variables were used, the only difference was that the team set the intercept equal to 0. Although this adjustment may slightly impact the accuracy of the model, the team believes it allows for a better interpretation of the input variables on Total streams. The overall model is again statistically significant, only this time the adjusted R^2 has increased to 44.1%. This means the overall model is a much better predictor of total streams and one can see that now the coefficients react in both the positive and negative direction.

Linear Regression Model 3: Total Streams as a result of BPM, Energy, Acousticness, Speechiness but the intercept is set to 0

```
Coefficients:
                 Estimate Std. Error t value Pr(>|t|)    
Beats_per_Minute  2088653     533850   3.912 9.79e-05 ***
Energy            3821014     898008   4.255 2.30e-05 ***
Acousticness      1855692     720354   2.576  0.01014 *  
Speechiness      -5089073    1854114  -2.745  0.00617 ** 
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 571900000 on 949 degrees of freedom
Multiple R-squared:  0.4429,    Adjusted R-squared:  0.4406 
F-statistic: 188.6 on 4 and 949 DF,  p-value: < 2.2e-16
```

In this third regression model, the team only looked at the focus variables and how those would impact total streams. Again, the overall model is significant and the adjusted R^2 remained constant at 44.1%. This model gives the team the below equation which can be used to help predict Total streams:

Total Streams = β1BPM + β2Energy + β3Acousticness + β4Speechiness
Total Streams = 2,088,653* BPM + 3,821,014 * Energy + 1,855,692* Acousticness – 5,089,073 * Speechiness

Linear Regression Model 4: Total Streams as a result of BPM, Energy, Acousticness, Speechiness but the intercept is set to 0 and the team took the log of Total Streams

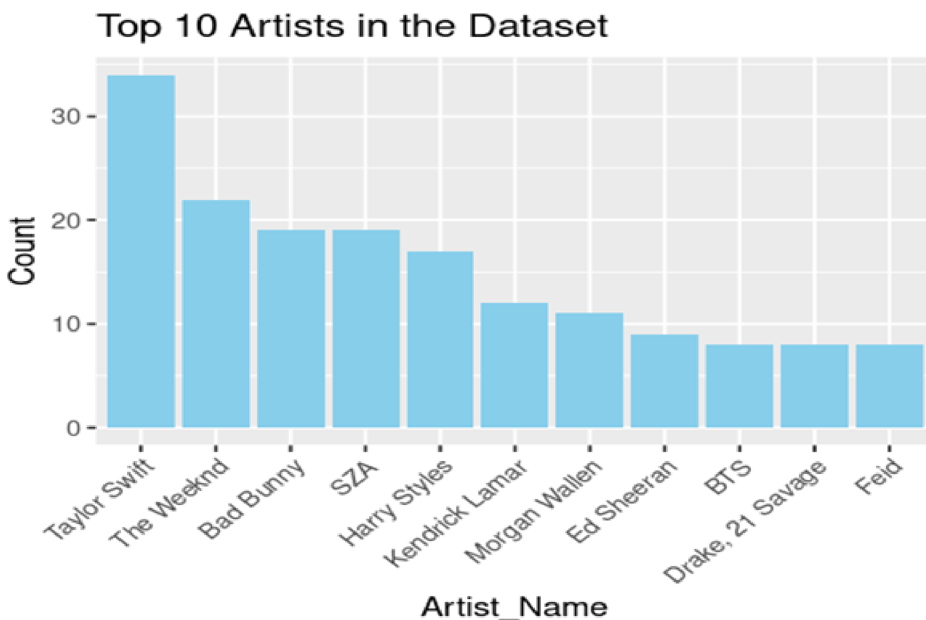
```
Coefficients:
                 Estimate Std. Error t value Pr(>|t|)    
Beats_per_Minute 0.059056   0.002853  20.697  < 2e-16 ***
Energy           0.150128   0.004800  31.278  < 2e-16 ***
Acousticness     0.071309   0.003850  18.521  < 2e-16 ***
Speechiness      0.028229   0.009910   2.849  0.00449 ** 
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.057 on 949 degrees of freedom
Multiple R-squared:  0.9756,    Adjusted R-squared:  0.9755 
F-statistic:  9487 on 4 and 949 DF,  p-value: < 2.2e-16
```

Because the input variables are percentages, the team determined that taking the log of total streams made sense to be able to better interpret the percent impact each input variable has on Total Streams. The result was a significant model and an R^2 of 97.6%. The coefficients can be interpreted as the percent increase in the dependent variable for every 1% increase in the independent variable. For example, a 1% increase in Energy will lead to a 15% increase in Total Streams.

After reviewing total streams for Spotify, the team reviewed the top charts. The same variables showed significance in the top charts list. This was compared to Apple and Deezer to see if the factors were the same for other top streaming services. Apple Chart hits are very close to Spotify Chart hits. Spotify top charts did not trend and show much correlation with Deezer top charts.

4.5 Top Songs and Artists

The below bar chart includes the top 10 Artists based on song count within the Spotify top 1,000 songs for 2023. Taylor swift unsurprisingly took the top spot with the announcement and start to her World Eras tour and her recent love interest with NFL star Travis Kelce. Other top Artists include The Weekend, Bad Bunny and SZA.



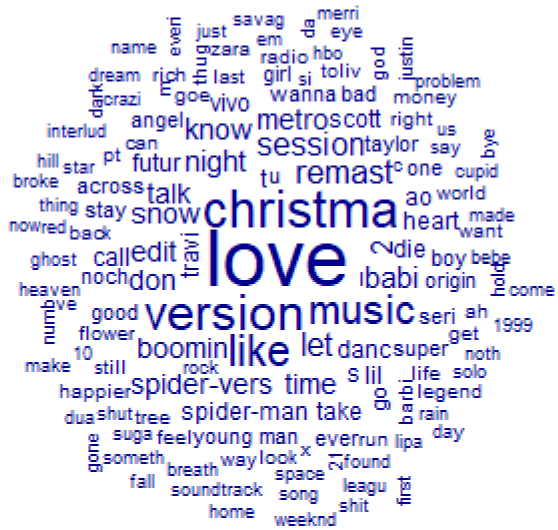
When you reconfigure that date to look at top Artists based on Total Streams (rather than song count), the Artists included in the top 10 change, but Taylor Swift and The Weekend still remain in the top 2 spots with The Weekend having ~132 million more total streams than Taylor.

	Artist_Name	Total_Streams
1	The Weeknd	14185552870
2	Taylor Swift	14053658300
3	Ed Sheeran	13908947204
4	Harry Styles	11608645649
5	Bad Bunny	9997799607
6	Olivia Rodrigo	7442148916
7	Eminem	6183805596
8	Bruno Mars	5846920599
9	Arctic Monkeys	5569806731
10	Imagine Dragons	5272484650

Both Taylor Swift and The Weekend are incredibly successful and although both considered pop, Taylor Swift is pure Pop while The Weekend is very R&B.

4.6 Text Mining

Produce that Beat thought it would be useful to analyze the song titles of the top 1,000 Spotify songs to determine if there were any key words that appeared multiple times among the top songs. The team leveraged text mining techniques to create a word cloud which not only included all words in song titles across the 1,000 songs, but also enlarged the words that appear most frequently. Stop words for English were initially removed and the output indicated that there were still several basic Spanish words such as "la" and "un". These were removed to get a better understanding of the title themes. Additionally, the team excluded music related words such as "Explicit", "Remastered", "Version", "Remix", and "Feat". Key themes include relationships with "love" and "like", holidays like Christmas, financial concerns including "broke" and "money", and religion including "God" and "angel". The team was able to pull a list of the top 10 words.



4.7 Max Spotify Counts

Max Total Streams: "*Blinding Lights"* by The Weekend

- BPM: 171
- Energy: 80%
- Speechiness: 7%
- Acousticness: 0%

Max Spotify Charts Count:" *Seven"* by Latto and Jung Kook

- BPM: 125
- Energy: 83%
- Speechiness: 4%
- Acousticness: 31%

Max Spotify Playlist Counts: "*Get Lucky*" by Pharrell Williams

- BPM: 116
- Energy: 81%
- Speechiness: 4%
- Acousticness: 4%

Taking the mean of the above metrics results in a song that contains 137 beats per minute and has Energy, Speechiness, and Acousticness levels of 81%, 5% and 12% respectively. The team ran a prediction model. The predicted output of stream values based on those metrics showed a value of ~593 million compared to the current max stream value of ~3.7 billion (Blinding Lights by The Weekend).

5.1 Data Science Questions

How can song characteristics be used to predict the success of future songs?

Predicting the success of future songs based on their characteristics is a complex task that often involves a combination of musical, cultural, and marketing factors. While guaranteeing success is challenging, specific characteristics have been identified as potential indicators. Here are some factors that are often considered:

- Streaming metrics such as playlist and chart counts
- Quantifying musical elements that are part of the top hits. Data showed that Energy, BPM, Acousticness, and Speechiness contribute to either being a top chart song or a popular song in a playlist
- Meaningful and catchy lyrics and melodies tend to be a more significant part of Spotify's top 1,000 songs
- Data showed a majority of songs were played in mode Major meaning that listeners prefer a happy and positive song

6.1 Future Considerations

The Spotify data set was instrumental in Produce that Beat's analysis and strategic pivot. The data analysis team did, however, uncover gaps in the data and missing data points which would have provided additional input and value to the analysis. For future analysis, the Label would look to obtain additional data to help drive greater insights and decision making. Some go forward considerations include:

- Artist and Streamer (user) Nationality (country of origin): The Label would like to understand if there is a correlation between the nationality of an artist and both the geographic breakdown of song streams and the nationality of the songs listeners
- Deezer is a streaming platform founded in France. The Label would like insight into Deezer's user demographics to understand if there is a correlation between country of origin and both the nationality of artists and the platform's users
- Song and Artist Genre: Understanding song and artist genres would provide insight into listeners' music preferences. Not only would it be useful in identifying the variability in user preferences, but it would allow Produce the Beat to predict what type of music listeners enjoy based on past streams
- Full Scope of Playlist Songs: Insight into all songs within a specific playlist would allow Produce that Beat to continue to gain insight into listener preferences and habits. Although the team was provided with how many playlists the songs were in, understanding what other songs were in the same playlist would be pivotal into predicting future trends and using generative AI to make recommendations

- Collaborations: Including the Artist Count data to understand the impact of featured artists on songs. The team would like to take a closer look into the collaboration aspect of song production
- Song Language: The analysis uncovered that many songs appeared to be in Spanish based on both the Song Title and Artist Name. The team would like to take a closer review of Spanish songs. This may have been overlooked and would be imperative in both the talent discovery and the music production process

7.1 Recommendations

Produce that Beat is excited to announce the next phase of its strategic pivot to becoming a leader within the music industry. The data analysis team has shared remarkable findings that the Label believes will position itself for immediate success and allow itself to not only compete for top artists, but consistently sign new contracts and produce chart topping hits.

The Label fosters and encourages creativity among its artists, and has determined that songs with the below metrics often times have greater odds of gaining widespread popularity. With that said, the Label is going to strongly consider the below metrics when working with its artists on their next album or newest single.

- BPM: 120-130
- Energy: 70-85
- Speechiness: 5-10
- Acousticness: 0-5

Second Produce that Beat is excited to announce the release of its very first Christmas album which is set to be released on December 20th. Given the holiday spirit, the peppiness of holiday music, and the frequency of words such as "Love" and "Christmas" in popular song titles, the Label believes that releasing a Christmas album will help carry success into 2024.

Lastly, the Label is working on its next hot single, "Love Session", an upbeat song that highlights all the wonderful aspects of being in love. This song title was strategically selected based on findings from leveraging text mining and word cloud analysis. The Label believes that titling the song "Love Session" and prescribing it to the above metrics will lead to the greatest chance of rocketing up the charts.

8.1 Conclusion

The strategic integration of Spotify data analysis into Produce that Beat's business strategy marks a pivotal moment in the label's evolution. Recognizing the need for innovation in talent discovery and music production, the Label is striving to combine industry expertise with cutting-edge analytics.

The two-fold data strategy to identify emerging artists and produce chart-topping songs demonstrates a commitment to staying ahead of industry trends and fostering a dynamic environment for talent growth. By leveraging Spotify data, Produce that Beat not only seeks to streamline the artist discovery process, but also aims to enhance its music production capabilities, ensuring a consistent release of chart-topping hits.

This initiative positions Produce that Beat at the forefront of the music industry's transformation, redefining how record labels navigate the complexities of talent discovery and song production. As the Label integrates insights from Spotify data with its historical industry experience, it is poised to not only compete effectively but also set new standards for success in an ever-evolving musical landscape.

Produce that Beat's commitment to innovation emphasizes its dedication to remaining a leader in the music industry. The Label's future success depends on the effective execution of this data strategy, aligning its operations with the ever-changing preferences of audiences and propelling the Label into a new era of talent discovery and music excellence.

9.1 References

Elgiriyewithana, Nidula. "Kaggle." *Most Streamed Spotify Songs 2023*, 26 Aug. 2023, www.kaggle.com/datasets/nelgiriyewithana/top-spotify-songs-2023/data. Accessed 17 Sept. 2023.

Saltz, Jeffrey S., & Stanton, Jeffrey M. (2021). *Data science for business with R*. SAGE Publications.