Spotify Tracks Dataset

A DATS 6103 Project by Team -3

By Aditya Nayak, Alexander Khater, Pooja Chandrashekara & Vaishnavi Nagarajaiah

Introduction

Music is something that we all enjoy during our lives. It is the sound that is brought together through a harmony of instruments and singing voices. The music industry we see today was not so smooth and easygoing in the past, Spotify brought a revolutionary change in the music industry satisfying the needs of both the audience and the music industry with its innovative tactics. It One of the most influential and used apps for audio streaming. Our problem statement is to identify music trends and characteristics in Spotify dataset extracted from Kaggle. We aim to look into the relationship between traditional music theory measurements such as key, tempo etc. Thus, we arrive at the question of what makes a song popular, our project aims to explore various factors that affect popularity of a song.

Content

This is a dataset of Spotify tracks over a range of 125 different genres. Each track has some audio features associated with it. The data is in CSV format which is tabular and can be loaded quickly.

Usage

The dataset is being used for observing the relationship between traditional music theory metrics (Key, Tempo, Time Signature, Duration, Tempo, Energy, Explicit, Mode) and a song's streaming popularity on Spotify and find out which factors contribute more towards a song be more popular.

Dataset Columns Description:

- **track_id:** The Spotify ID for the track
- **artists:** The artists' names who performed the track. If there is more than one artist, they are separated by a;
- **album_name:** The album name in which the track appears.
- **track_name:** Name of the track.
- **popularity:** The popularity of a track is a value between 0 and 100, with 100 being the most popular. The popularity is calculated by algorithm and is based, in the most part, on the total number of plays the track has had and how recent those plays are. Songs that are being played a lot now will have a higher popularity than songs that were played a lot in the past. Duplicate tracks (e.g. the same track from a single and

an album) are rated independently. Artist and album popularity is derived mathematically from track popularity.

- **duration_ms:** The track length in milliseconds.
- **explicit:** Whether the track has explicit lyrics (true = yes it does; false = no it does not OR unknown).
- **danceability:** Danceability describes how suitable a track is for dancing based on a combination of musical elements including tempo, rhythm stability, beat strength, and overall regularity. A value of 0.0 is least danceable and 1.0 is most danceable.
- **energy:** Energy is a measure from 0.0 to 1.0 and represents a perceptual measure of intensity and activity. Typically, energetic tracks feel fast, loud, and noisy. For example, death metal has high energy, while a Bach prelude scores low on the scale.
- **key:** The key the track is in. Integers map to pitches using standard Pitch Class notation. E.g., 0 = C, 1 = C♯/D♭, 2 = D, and so on. If no key was detected, the value is -1.
- **loudness:** The overall loudness of a track in decibels (dB).
- **mode:** Mode indicates the modality (major or minor) of a track, the type of scale from which its melodic content is derived. Major is represented by 1 and minor is 0
- **speechiness**: Speechiness detects the presence of spoken words in a track. The more exclusively speech-like the recording (e.g., talk show, audio book, poetry), the closer to 1.0 the attribute value. Values above 0.66 describe tracks that are probably made entirely of spoken words. Values between 0.33 and 0.66 describe tracks that may contain both music and speech, either in sections or layered, including such cases as rap music. Values below 0.33 most likely represent music and other non-speech-like tracks.
- **acousticness:** A confidence measure from 0.0 to 1.0 of whether the track is acoustic. 1.0 represents high confidence the track is acoustic.
- **instrumentalness:** Predicts whether a track contains no vocals. "Ooh" and "aah" sounds are treated as instrumental in this context. Rap or spoken word tracks are clearly "vocal". The closer the instrumentalness value is to 1.0, the greater likelihood the track contains no vocal content.
- **liveness:** Detects the presence of an audience in the recording. Higher liveness values represent an increased probability that the track was performed live. A value above 0.8 provides a strong likelihood that the track is live.
- **valence:** A measure from 0.0 to 1.0 describing the musical positiveness conveyed by a track. Tracks with high valence sound more positive (e.g., happy, cheerful, euphoric), while tracks with low valence sound more negative (e.g., sad, depressed, angry).
- **tempo:** The overall estimated tempo of a track in beats per minute (BPM). In musical terminology, tempo is the speed or pace of a given piece and derives directly from the average beat duration.
- **time_signature:** An estimated time signature. The time signature (meter) is a notational convention to specify how many beats are in each bar (or measure). The time signature ranges from 3 to 7 indicating time signatures of 3/4, to 7/4.

- **track_genre:** The genre in which the track belongs.

Overview and SMART Questions:

We choose this dataset mainly because we are all avid music listeners and one of us has subject matter expertise in the field of music creation and this dataset offered potential insight into certain questions. Since we had interest in the field of making music, we wondered, "What makes a song popular?". We wanted to pursue the hypothetical "magic song"; a song that will become popular purely based on its base musical qualities. This brought us to our first of 3 SMART questions:

1. **"Can we use regression modeling on musical data about a given song to predict its popularity?"**

Our Subject Matter Expert did not think our models would be that successful because popularity in music, as shown in other similar studies, has more to do with the musical and cultural landscape around it as opposed to the song itself. It makes intuitive sense, imagine a popular song from today being released in 1955. The cultural attitudes towards music would not be ready for it yet and it would probably not be very popular. Yet because we were deeply curious, we still wanted to see how well we could predict popularity based on a song's musical qualities.

During dataset analysis, our team Subject Matter Expert noticed that there were two kinds of variables that were used to describe music: traditional, and algorithmic. Meaning some of the variables such as key, mode, tempo, time signature, explicitness, and duration are classical metrics for music (or in the case of explicitness, has become widely accepted over the last forty years. These metrics are widely understood with exact definitions in the realm of music. The algorithmic metrics, such as danceability, valence, energy, speechiness, etc. reference qualities that are often talked about but have no exact definition. This means that these values in the data were completely assigned by Spotify's algorithm and are beholden only to Spotify's definitions, rather than the musical community. This creates an interesting inquiry of whether Spotify's metrics for describing music are more effective for predicting popularity than traditional metrics. This leads to our second SMART question:

2. **"Do Regression Models predicting popularity do better using traditional music metrics as predictor variables or Spotify's Algorithmic Musical Metrics?"**

Our Subject Matter Expert did not know whether the Spotify Metric Model would outperform the Traditional Metric Models, but there is a reason to believe the biggest music streaming company would develop effective metrics to describe streamed music.

Our third SMART question started more as a side inquiry and became a larger project as we started to see success. We were curious about what would make a song positive and danceable per Spotify's algorithmic "valence" and "danceability" metrics. Would traditional or algorithmic metrics do better? We made our third SMART question about predicting the combination variable of valence and danceability:

3. **"Can we predict what makes a positive danceable song using regression modeling? Will it be best predicted by traditional metrics, algorithmic metrics, or a combination of the two?"**

We also did some modeling of the "durability variable", but that was more out of pure curiosity and did not have any SMART Question driving it.

Data Manipulation

1. **Cleaning the Data:**

This is a very clean dataset with no null values, so no null value handling was required.

FOR THE MODELING PHASE:

2. **Dropping Non-Interesting Variables:**

Since our SMART question was more focused on the inherent qualities of the music in regression, before building our models, we dropped all non-numerical data: such as 'track_id', 'artists', 'album_name', and 'track_name'

Note: The artist variable would have been interesting to use, but sadly it was coded in a way that listed artist collaborations as separate artists and had no way to account for different artists with the same name. This would have required a lot of work to clean into a usable state that was outside the scope of our initial SMART question but would be great for future inquiry.

3. **Encoding:**

We encoded the binary variables as float type 0 or 1 variables, for regression and upcoming VIF tests

4. **Multicollinearity:**

Because of the results of our correlation matrix, as you'll see below, we decided to run a VIF test on the data before modeling. After running a VIF test on our dataset we found deep multicollinearity concerns with VIF scores exceeding 50. In order to control this, we took 3 steps to strip the concerns from our data:

i. We removed 'loudness' and 'energy' as they were both too heavily correlated with many of the other variables

ii. We merged 'Valence' and 'Danceability' as they were heavily correlated with each other and described similar concepts. The new variable is simply called ('valence+danceability).

iii. We centered our data by subtracting the mean from each one of our variables. This handled any structural multicollinearity in our new combination variables

After these changes, all our variables showed satisfactory VIF scores, and we were ready to model.

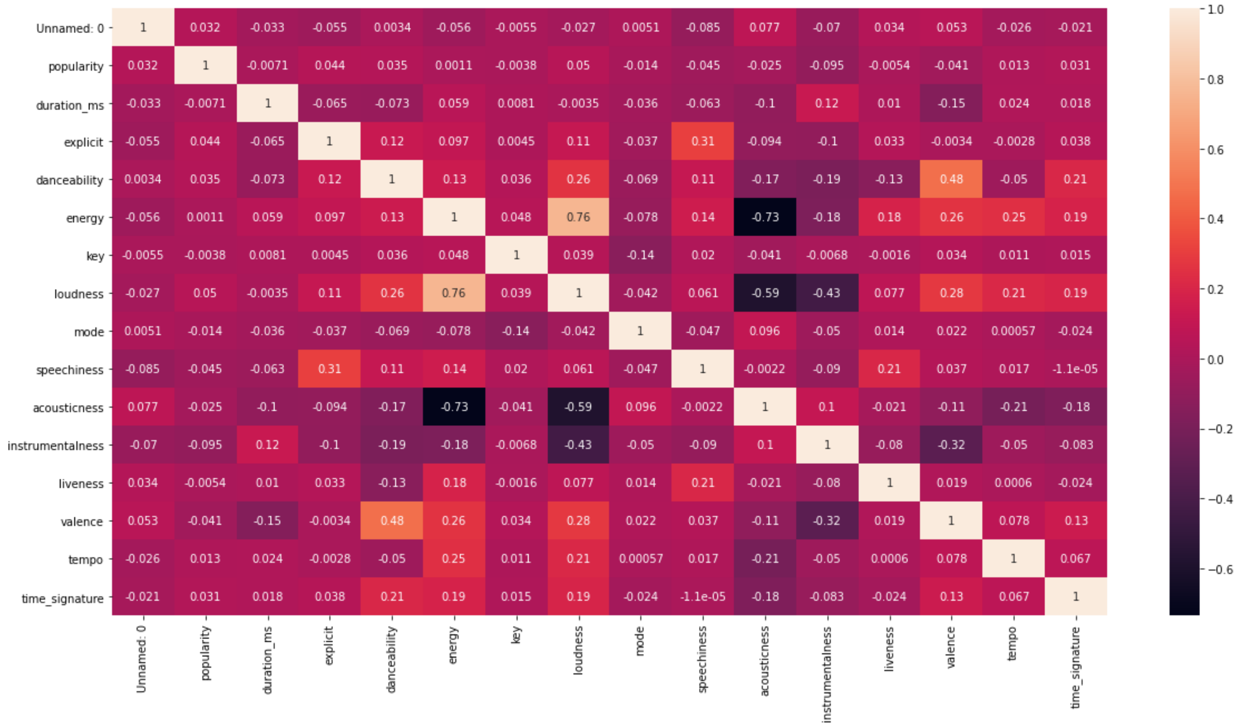
EDA Summary

Exploratory data analysis (EDA) is a vital phase in any data analysis process. In order to direct particular testing of your hypothesis, the primary goal of the exploratory analysis is to look at the data for distribution, outliers, and anomalies. It also offers aids for developing hypotheses by helping people see and comprehend the data, typically through graphical representation.
Here, in our project, we performed some standard EDA techniques such as univariate analysis, Bi-Variate analysis, and multivariate analysis.

Univariate analysis: Univariate Analysis is the simplest form of statistical analysis that is performed on one variable. The key variable on which the analysis is performed decides which statistical techniques can be used such as line plot, histogram, boxplot, etc.
In our project, we have used boxplot analysis popularity, danceability, energy, acousticsness, and valence. Similarly, we used a count plot for key and mode features. The univariate analysis of the above said variables gave the following outcomes.

- Majority of songs in the dataset are mostly unpopular.
- Majority of songs are danceable.
- Majority of tracks have high energy.
- All genres are equally divided in the dataset
- Majority of tracks have key-7 i.e., G pitch and key-0 i.e., C pitch in them.
- Majority of songs have major code modality in them.
- Majority of songs have low acoustics.
- Many songs seem to have a neutral emotional outcome in them.

Bi-variate analysis: One of the simplest types of quantitative analysis is bivariate analysis. In order to understand their empirical link, it involves the analysis of two variables.

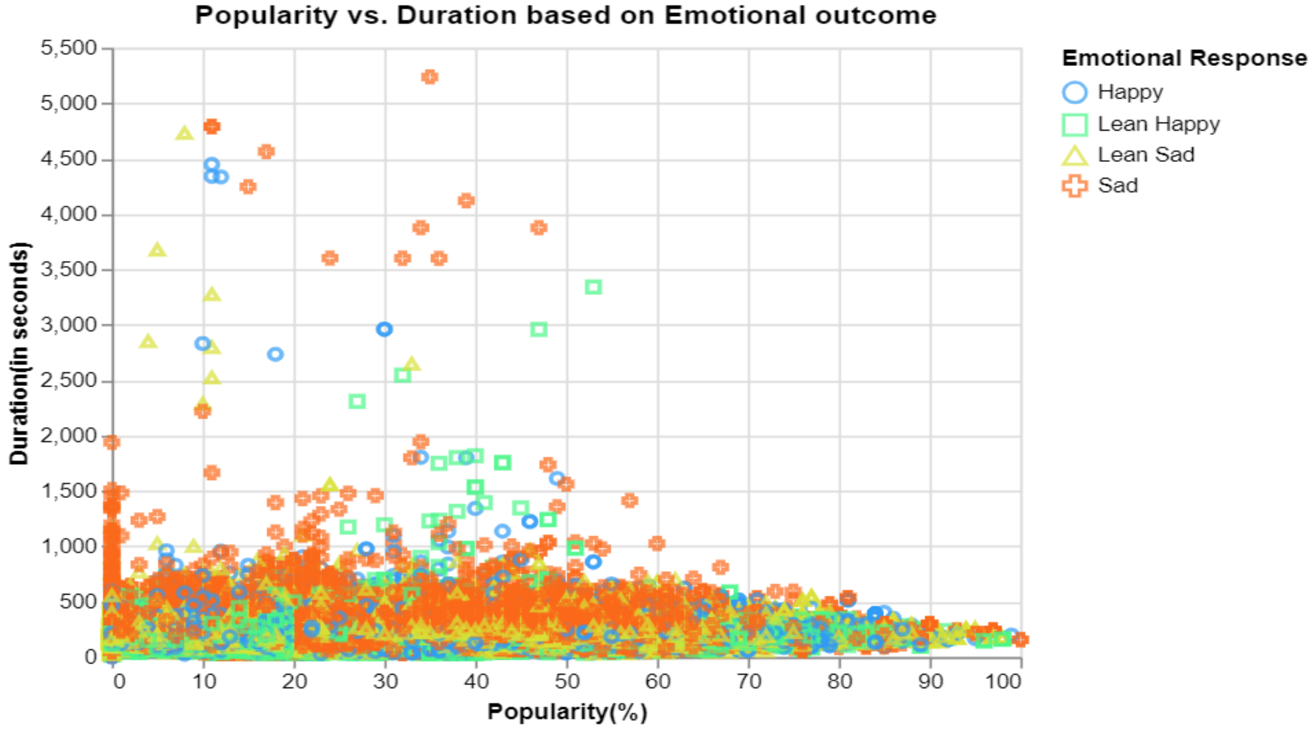


	Unnamed: 0	popularity	duration_ms	explicit	danceability	energy	key	loudness	mode	speechiness	acousticness	instrumentalness	liveness	valence	tempo	time_signature
Unnamed: 0	1	0.032	-0.033	-0.055	0.0034	-0.056	-0.0055	-0.027	0.0051	-0.085	0.077	-0.07	0.034	0.053	-0.026	-0.021
popularity	0.032	1	-0.0071	0.044	0.035	0.0011	-0.0038	0.05	-0.014	-0.045	-0.025	-0.095	-0.0054	-0.041	0.013	0.031
duration_ms	-0.033	-0.0071	1	-0.065	-0.073	0.059	0.0081	-0.0035	-0.036	-0.063	-0.1	0.12	0.01	-0.15	0.024	0.018
explicit	-0.055	0.044	-0.065	1	0.12	0.097	0.0045	0.11	-0.037	0.31	-0.094	-0.1	0.033	-0.0034	-0.0028	0.038
danceability	0.0034	0.035	-0.073	0.12	1	0.13	0.036	0.26	-0.069	0.11	-0.17	-0.19	-0.13	0.48	-0.05	0.21
energy	-0.056	0.0011	0.059	0.097	0.13	1	0.048	0.76	-0.078	0.14	-0.73	-0.18	0.18	0.26	0.25	0.19
key	-0.0055	-0.0038	0.0081	0.0045	0.036	0.048	1	0.039	-0.14	0.02	-0.041	-0.0068	-0.0016	0.034	0.011	0.015
loudness	-0.027	0.05	-0.0035	0.11	0.26	0.76	0.039	1	-0.042	0.061	-0.59	-0.43	0.077	0.28	0.21	0.19
mode	0.0051	-0.014	-0.036	-0.037	-0.069	-0.078	-0.14	-0.042	1	-0.047	0.096	-0.05	0.014	0.022	0.00057	-0.024
speechiness	-0.085	-0.045	-0.063	0.31	0.11	0.14	0.02	0.061	-0.047	1	-0.0022	-0.09	0.21	0.037	0.017	-1.1e-05
acousticness	0.077	-0.025	-0.1	-0.094	-0.17	-0.73	-0.041	-0.59	0.096	-0.0022	1	0.1	-0.021	-0.11	-0.21	-0.18
instrumentalness	-0.07	-0.095	0.12	-0.1	-0.19	-0.18	-0.0068	-0.43	-0.05	-0.09	0.1	1	-0.08	-0.32	-0.05	-0.083
liveness	0.034	-0.0054	0.01	0.033	-0.13	0.18	-0.0016	0.077	0.014	0.21	-0.021	-0.08	1	0.019	0.0006	-0.024
valence	0.053	-0.041	-0.15	-0.0034	0.48	0.26	0.034	0.28	0.022	0.037	-0.11	-0.32	0.019	1	0.078	0.13
tempo	-0.026	0.013	0.024	-0.0028	-0.05	0.25	0.011	0.21	0.00057	0.017	-0.21	-0.05	0.0006	0.078	1	0.067
time_signature	-0.021	0.031	0.018	0.038	0.21	0.19	0.015	0.19	-0.024	-1.1e-05	-0.18	-0.083	-0.024	0.13	0.067	1

- As the danceability of the song increases, the positiveness conveyed by the track also increases.
- As the energy of the song increases, the loudness of the song also increases.
- As the energy of the song increases, the acoustics of the song increase.
- As acoustics increases, the loudness of the song decreases therefore the energy of the song also decreases.
- As the loudness of the track increases, the instrumentals used in the song decrease hence the energy also decreases.
- As instrumentals increase, the positiveness conveyed by the track decreases hence the danceability also decreases.

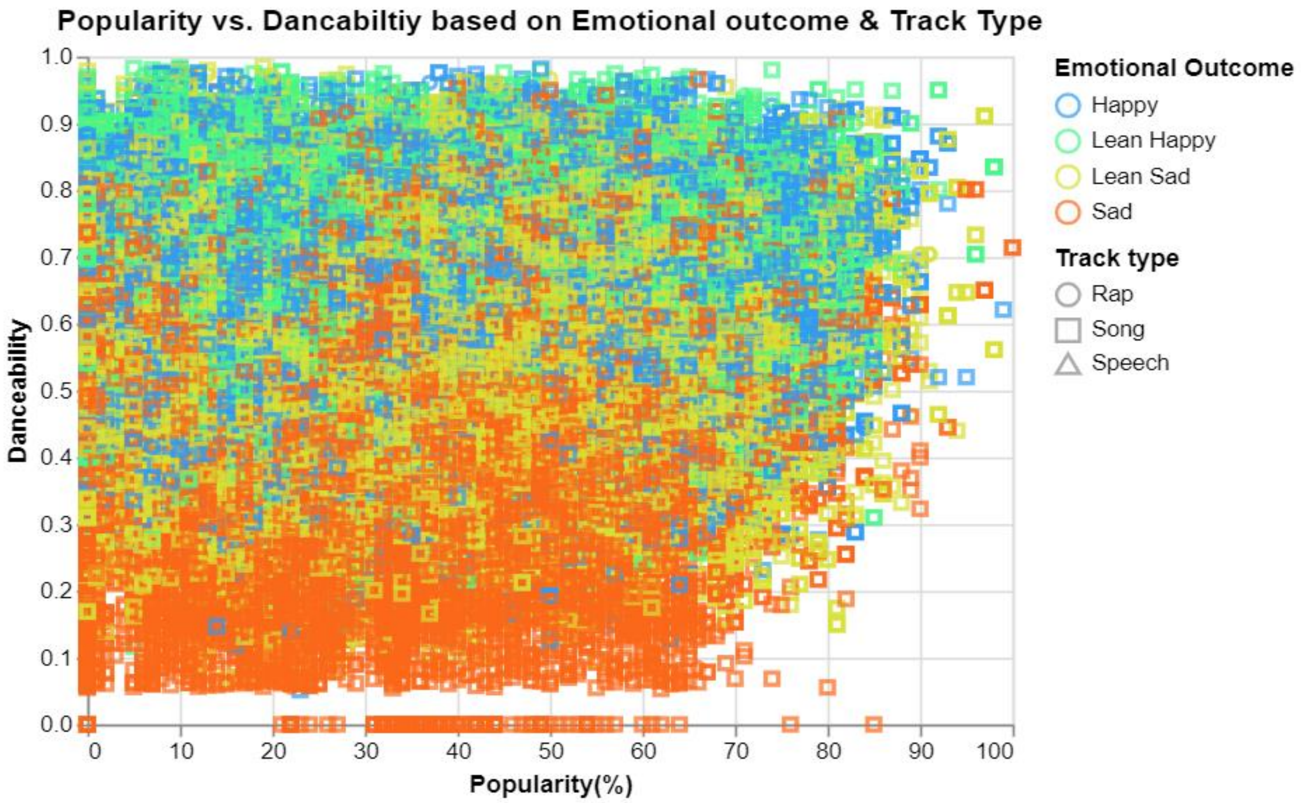
Multi-variate Analysis: The goal of multivariate analysis (MVA) is to find any potential associations between numerous variables (more than two). It helps us in analyzing all the possible independent variables and their relationship with each other.

In our project, we have done two multi-variate analyses. One is with population vs duration of a song based on the emotional outcome, second is the analysis of population vs danceability based on emotional outcome and track genre. The figures below show the correlation scatter plot of the variables.



The above plot led us to make the following conclusions.

- Most tracks are between 0 to 1000 seconds.
- The most popular song is in less than 500 seconds.
- Few unpopular or least popular songs are of high duration.
- Majority of unpopular and average popular songs exhibit sad emotions.
- The most popular songs are likely to exhibit happy emotions.
- The most popular song is of very less duration and sad song.
- The lean sad songs are comparatively less in the dataset.



The above multi-variate plot led us to the following conclusions:

- The least danceable songs exhibit sad emotions.
- There are only a few raps and speech-type tracks in the dataset.
- Majority of songs exhibit sad emotions.
- The average danceable songs and average popular songs exhibit lean sad emotion.
- Few of the most popular and danceable songs exhibit sad emotions.
- Highly danceable songs exhibit happy and lean happy emotions.
- Highly danceable and highly popular songs exhibit happy emotion.

Statistical Modelling

Regression models are used to predict a continuous outcome variable based on one or more predictor variables. There are several types of regression models that we have considered while working on our project, including multiple linear regression, polynomial regression, KNN regression and XGBoost regression. Each type of model is suited for different types of data and different types of prediction problems.

One of the main advantages of using regression models is that they can be used to identify the strength of the relationship between the predictor variables and the outcome variable, and they can also be used to estimate the importance of each predictor variable in predicting the outcome. This can clarify the underlying patterns and trends in the data, and in making informed decisions based on the insights gained from the model.

➢ Models for predicting popularity of an audio track:

To predict the popularity of the track song, as stated before, we noticed that there are variables in the dataset that use traditional and other metrics that are not generally considered for describing an audio type and decided to divided all the parameters into two parts: 1) Traditional terms which are quite familiar in the music domain such as tempo, key, mode etc. and 2) Algorithmic variables which are not usually used in music terms such as valence, danceability, speechiness etc.

1. **Multiple Linear Regression**
   To understand if there is any relationship with the traditional and algorithmic variables for determining popularity of an audio track, we first build the simple multi-variable linear regression models. On passing the data, we notice that the model predicts the popularity accurately by only *0.31%* using traditional terms and *1.4%* using algorithmic variables. As the accuracy of these models is very low, they are not a good fit for our data.

2. **Polynomial Regression**
   Another model that we decided to build was the Polynomial regression model. A type of regression analysis in which the relationship between the independent variable x and the dependent variable y is modeled as an nth degree polynomial. Polynomial regression can be used to model relationships between variables that are not linear. On passing the data we observe that the model can determine the popularity with an accuracy of 0.47% using traditional terms and 3.7% using algorithmic variables. These results are somewhat better than the accuracy obtained from the linear regression models. However, they still have a low accuracy score and hence is not a suitable model for our dataset.

3. **KNN Regression**
   This is the type of regression model that uses a set of labeled data points to make predictions about a continuous target variable. It works by finding the K nearest neighbors of a new data point, and then averaging their values to make a prediction. In our case, the ideal value of k is 5. On building this model, the models give a quite better accuracy of 34.55% when using the traditional metrics and 38.88% by using algorithmic terms which is still not good enough. Therefore, after building all these models, we can conclude that popularity is a variable that cannot be determined statistically using our given dataset.

➢ Models for predicting ‘valence+danceability’ of an audio track:

After concluding that popularity of an audio track cannot be determined, we decided to check whether we are able to predict if an audio track can be used to dance on or not. From our EDA, you can notice there is a gradation in the emotional outcome of audio tracks as danceability of a song increases in our Popularity vs. Danceability scatterplot. In addition to this, the terms valence and danceability have been merged due to their high correlation and possessing similar concepts.

1. **Multiple Linear Regression**
   On building our multiple linear regression models, we can note that the accuracy score of the model using traditional terms comes to *5.99%* using traditional terms and *12.27%* with the algorithmic variables. This score despite being low is somewhat better than the accuracy score obtained from the linear regression models built for predicting popularity. Hence, we can further investigate the variable.

2. **Polynomial Regression**
   On building our Polynomial regression models, we observe that the models give an accuracy score of *10.63%* using traditional and *26.95%* using algorithmic terms. The accuracy score is significantly better and from this we can form a hypothesis that the danceability and valence of a song can be predicted from the given data variables in our dataset.

3. **KNN Regression**
   To confirm our hypothesis, we can use the KNN regression model to determine if our hypothesis is true or not. On creating the models, they have an accuracy score of *30.54%* when using traditional and *49.29%* using algorithmic variables.

➢ Models for predicting the ‘danceability’ of an audio track:

Since the popularity of the audio track cannot be predicted and even the accuracy of the ‘valence+danceability’ prediction was low, we decided to perform the model testing on the variables combined with traditional and algorithmic approaches. So, we considered 'energy', 'instrumentalness', 'acousticness', 'liveness', 'loudness', 'tempo', and 'valence' features to predict the danceability of a track.

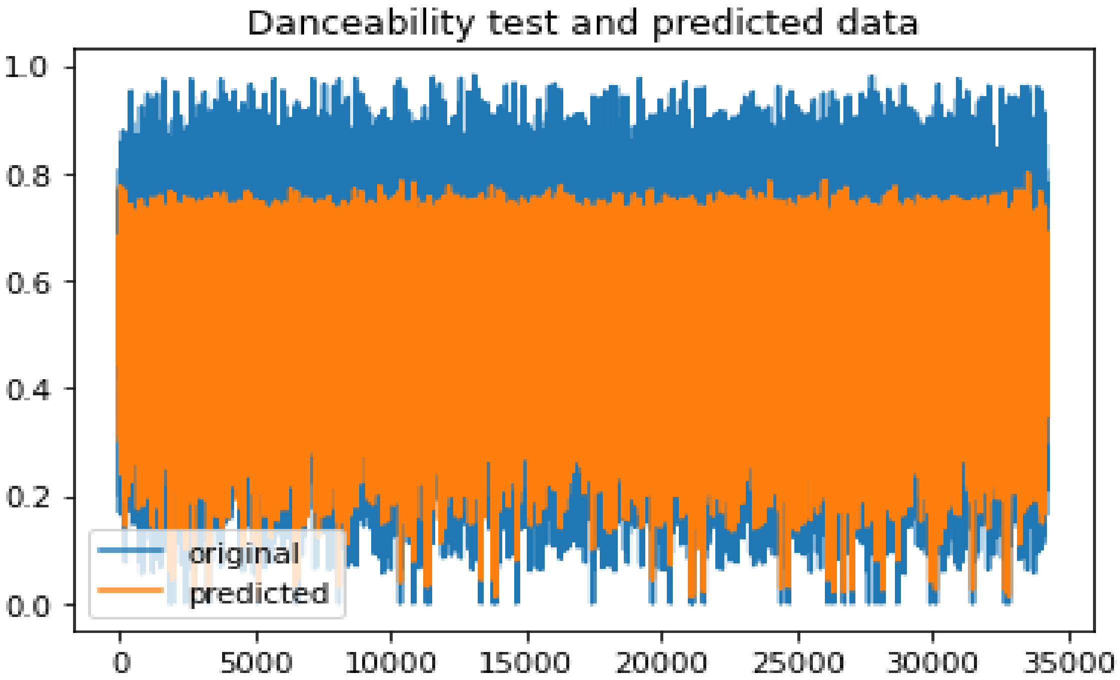
1. **Linear Regression**
   On building the linear regression model on the combined variables of traditional and algorithmic approaches, we got an accuracy of 30.93% in predicting the danceability of tracks in the dataset.

2. **KNN Regression**
   To check will there be further improvement in the accuracy of the model, we used K-Nearest-Neighbors model on the combined variables of traditional and algorithmic approaches. We got an accuracy of 41.65% for the KNN regression model in predicting the danceability.

3. **XGBoost Model**

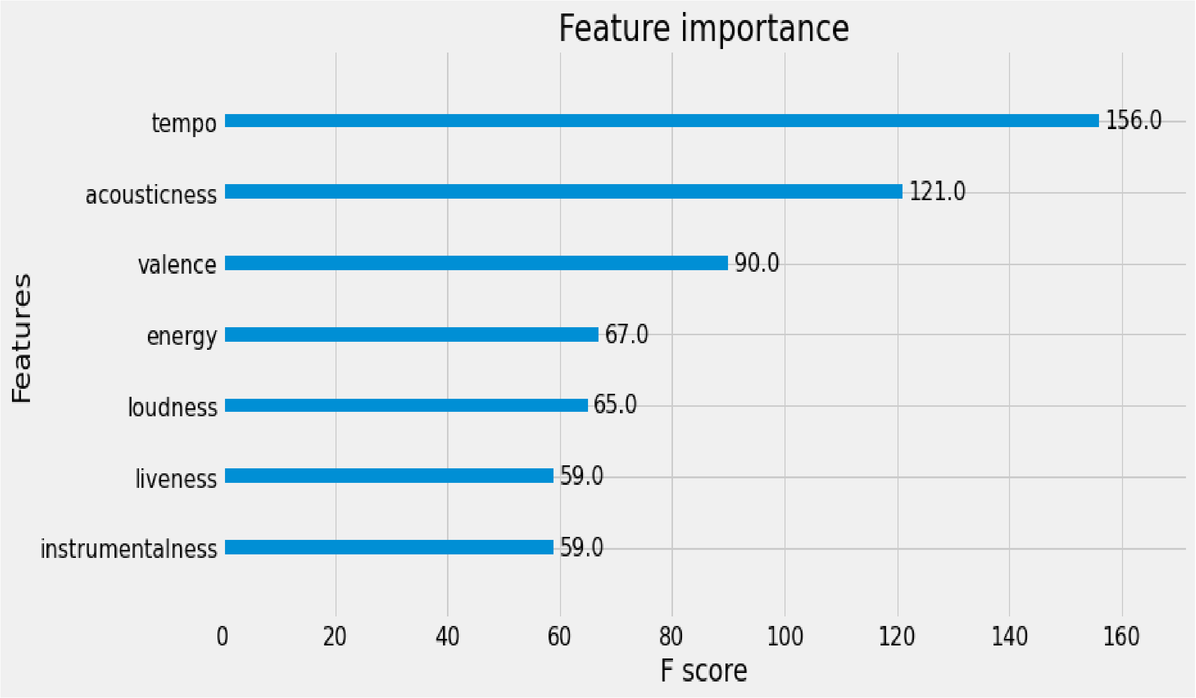
The Gradient Boosting Regression model determines the difference between the current prediction and the known correct target value. This variation is referred to as residual. After that, a weak model that maps features to that residual is trained using gradient boosting regression. This procedure moves the model closer to the desired outcome by adding the residual predicted by a weak model to the input of the current model. To further improve the accuracy, we used the XGBoost model, and we were able to achieve 52.23% accuracy.
The figure below shows the graph of the danceability test vs predicted data.

We can see that only half of the predicted value is being overlapped by the original value. This justifies the 52% accuracy we got.



Since we had taken the traditional variables, that doesn't exhibit any correlation with danceability. We wanted to know which feature is given the highest importance in the XGBoost model. The figure below shows the F-score of the features considered.

We can see that only half of the predicted value is being overlapped by the original value. This justifies the 52% accuracy we got.
Since we had taken the traditional variables, that doesn't exhibit any correlation with danceability. We wanted to know which feature is given the highest importance in the XGBoost model. The figure below shows the F-score of the features considered.



The F score is a measure based on the number of times a variable is selected for splitting, From the graph we can say that temp is given highest importance in the feature selection. Even though there is no correlation between tempo and danceability in the dataset we can see highest importance is given to tempo variable followed by acousticness, valence, and energy. This justifies that the traditional model combined with the algorithmic approach gives better results. From these findings, we can infer that Spotify has created new parameters that help in the data analysis and building prediction models.

Conclusion:

- Our problem statement was to identify music trends and characteristics in Spotify data and wanted to investigate relationship between traditional music metrics and Spotify automated metrics. We performed data cleanup and exploration and observed that our dataframe did not have any null values, therefore no observations were dropped, and no datatype conversions were required.
- After conducting the EDA analysis and attempting to determine what factors influence the popularity of songs in the dataset, we noticed that popularity had no correlation with any of the variables in the dataset. This led us to conclude that Spotify uses an automated method rather than a traditional approach to its dataset. It was also found that the majority of songs have high intensity and mode, low acoustics, and are danceable. All genres in the Spotify collection are evenly distributed, and most popular songs appear to have lean sad undertone.
- After performing modeling on popularity metrics using polynomial, multilinear and KNN regression methods. We learned that what makes a song popular is not inherent

in the song itself, and Spotify's techniques for identifying music in the age of streaming have improved.

- When we looked at the danceability metric for modeling, we discovered that traditional measures do not contribute much when compared to the algorithmic elements in all of the models. So, we thought of trying something different, therefore we considered combining traditional and algorithmic metrics because the algorithmic model gives the best results, but the traditional model can give better results with the help of the traditional model, and we performed the XGBoost modeling technique and obtained better accuracy.
- As a consequence of all the modeling methodologies, we decided on danceability as our goal variable because of its strong relationship with other parameters like valence, instrumentality, energy, loudness, and so on. We can also conclude that out of the three models we built XGBoost gave the best accuracy.

References:

1. Noah Askin and Michael Mauskapf. (June 2017). "What Makes Popular Culture Popular? Product Features and Optimal Differentiation in Music." Retrieved from: https://doi.org/10.1177/0003122417728662.

2. Cole, Tom. (October 2010). "You Ask, WE ANSWER: 'Parental Advisory' Labels - the Criteria and the History."
Retrieved from: https://www.npr.org/sections/therecord/2010/10/29/130905176/you-ask-we-answer-parental-advisory---why-when-how.