Final Project Report - Introduction to Data Science
Group members: Asmi Narsay, Mahek Kadakia, and Masumi Chhabria

This is a link to our dataset:
https://www.kaggle.com/datasets/solomonameh/spotify-music-dataset

The issue we are solving by analyzing this particular data set is that we can assist music artists with understanding current musical trends and music that is more likely to go viral. By doing so, music artists can use this information to create music within their genre that can reach as many people as possible, and help them increase their fanbase. Not only will this project's findings assist in helping artists learn how to better market their music, but it will also help them see what the newer generations are interested in listening to.

Some strategic aspects that we used in this project include Data Analysis and Data Visualization. More specifically, we analyzed how different factors in music, such as tempo, genre, danceability, etc, affected how many streams a song can get, on Spotify. We analyzed this data by creating multiple visualizations, such as histograms, bar plots, scatterplots, etc, to display correlations between these factors. By doing this, we were able to explain trends to music artists throughout many different industries. However, a very crucial note is that we must analyze these data visualizations and interpret them in many different ways because they can be viewed differently, from person to person.

Our project relates to many topics discussed in the lectures. Specifically, the emphasis on Data Visualization, as mentioned earlier in this segment. One of our main forms of Data Visualization comes from creating graphs and plots, to see how well the different factors in a song correspond to each other and the song's popularity. In addition to this, we also used NumPy and Pandas libraries, which are heavily used in discussions we have in class. By using these libraries, we were able to work with our dataset much more efficiently. For example, we used these libraries to find patterns within the data and present them in a comprehensible manner. We also use another topic discussed in class: Statistics. By using Statistics, we determined how strong correlations between variables are and even created regression models to test out the relationships between said variables. By using NumPy, Pandas, and Statistical concepts together, we have formed a proper analysis of our data.

Our project is a way to provide insight into recent patterns of music consumption. This database helps us to understand why certain music is more popular than others by analyzing various characteristics of songs. This study is important because it can be used in the music industry to predict future trends, which can be useful for artists, marketing, and user engagement. Artists and marketers can better tailor their work to suit the general public's taste, producing better results. It can also be utilized in large platforms like Spotify or Apple music to create a better recommendation system and increase usage. Music is all around us and we are constantly exposed to it, which is why we are excited to study more about the happenings behind it.

There have been past studies which utilize Spotify datasets, which address similar questions of user engagement, music popularity, and recommendations. For instance, there was a study done on Spotify podcast episodes that analyzed audio and transcripts to see what was the most popular and helped to create a more personalized page for podcast listeners. The music industry is ever growing and will continue to be a fundamental part of society and media, so gauging a strong understanding of what really makes it so popular is necessary to keep it modern and beneficial.

Our data is a structured numerical and categorical dataset in tabular format from kaggle, with 29 audio and descriptive features. The numerical features include energy and tempo, which represents the speed of the track in beats per minute (BPM). Another feature, loudness, measures the overall volume in decibels (dB). Other numerical attributes include liveness, which assesses the likelihood that the track was recorded live, and valence, which represents the positivity or happiness of the song.

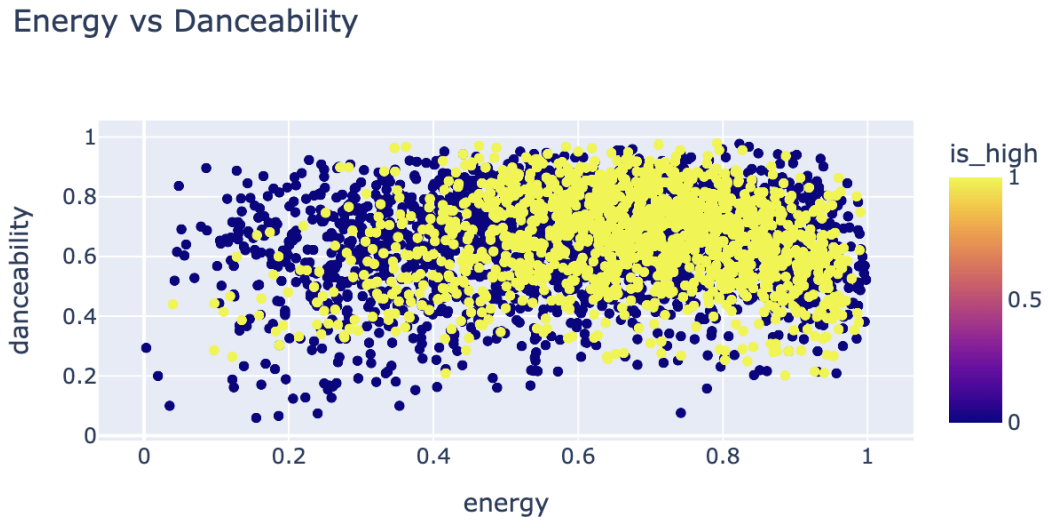
This dataset also includes categorical features such as playlist genre, which categorizes songs based on their genre, and track artist, identifying the performers. The dataset contains aggregated distributions of each feature, showing the number of songs that fall into specific value ranges and the numerical values for each feature, along with corresponding artist names and genre labels. Most importantly, it includes the track popularity which is calculated based on the total number of streams in relation to other songs which is our target variable for predictions.

We used supervised learning, a linear regression model and a random forest trees regression model, to determine which features of a song and what value of those features makes a song appealing and popular. A linear regression model helps to predict song popularity by identifying relationships between the song features and its popularity score. By ranking the regression coefficients from greatest to least, we were able to understand what features genuinely contribute to making a song catchy.

Since features in music often are interconnected, such as loudness and tempo, or energy and danceability, we also used Random Forest because it analyzes how song characteristics interact with each other and contribute to popularity. The model trained each tree on a random subset of music data and averaged its results to highlight any non-linear relationships and feature importance. We prepped the data by addressing any missing values and converting relevant categorical variables like playlist genre using label encoding. Based on learnings in class, we treated our training data and testing data separately, so using that we accordingly split the data with a 20% testing set and 80% training set. With suitable training, we then implemented the linear regression and random forest models to predict song popularity. We then tested the model with accuracy comparisons to find a percentage of how accurate the actual results were vs the predicted results with an R-2 score, mean absolute error, and mean squared error to see if our features are valid predictors of song popularity.

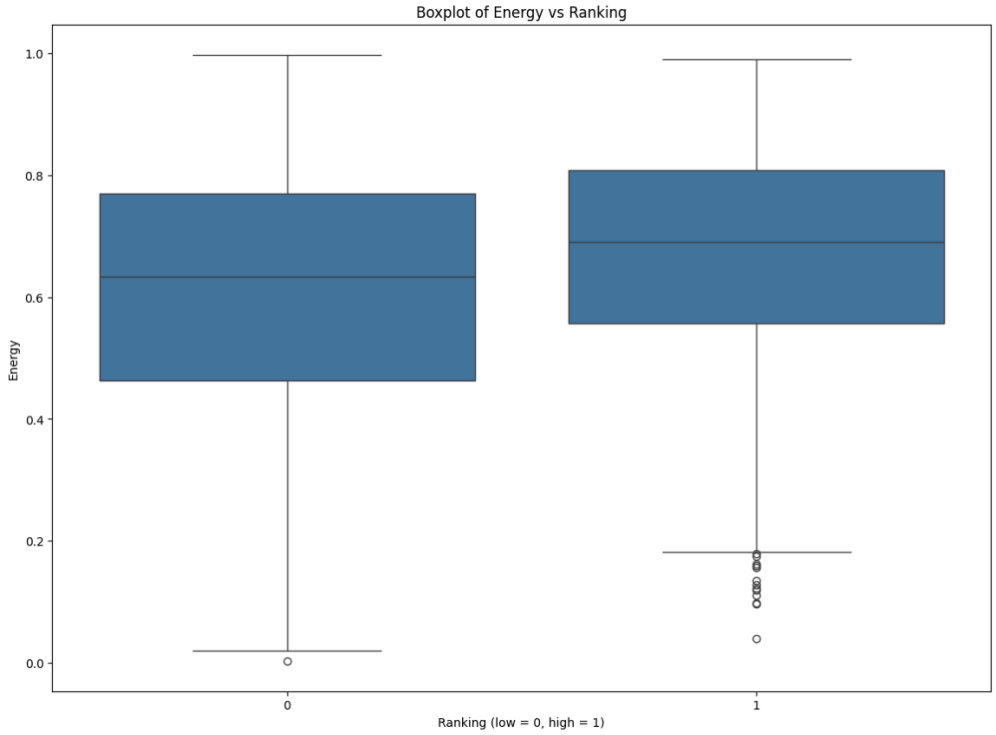
After outlining our project and applying all these steps, we are able to analyze our dataset efficiently. Let’s take a look at what results we gathered.

We created 5 suitable graphs that gave us important information about the different factors in our dataset. This first graph compares energy versus danceability using a scatter plot:



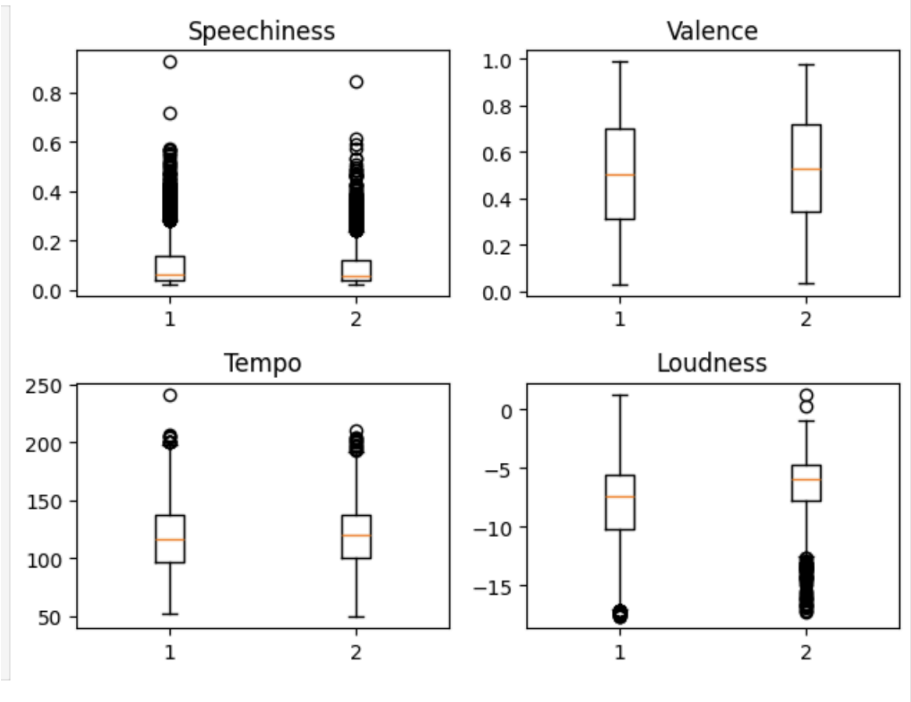
As you can see, the points are mostly clustered where danceability and energy are both relatively high. This means that both danceability and energy go hand-in-hand across all songs. However, it is important to acknowledge that although both low and high-ranking songs are clustered towards regions where energy is higher, the points representing high ranking songs are sparse where energy is low. This could indicate that although both low and high-ranking songs have a similar range in terms of danceability, lower-ranking songs are more likely to lack energy.

The next graph we looked is a boxplot of the distributions of genre on liveness:



The ranking with the smallest IQR range and variance is higher ranking music. This means that higher ranked music has relatively consistent energy levels while lower-ranked songs have varying energy levels. As seen in the figure, higher ranked music has a larger median energy level than lower-ranked music. It is very important to acknowledge that higher-ranked music has a smaller range of values which are on the higher energy side of the graph, with outliers on the lower energy level side. This means that it is more rare for higher-ranked songs to have lower energy levels.

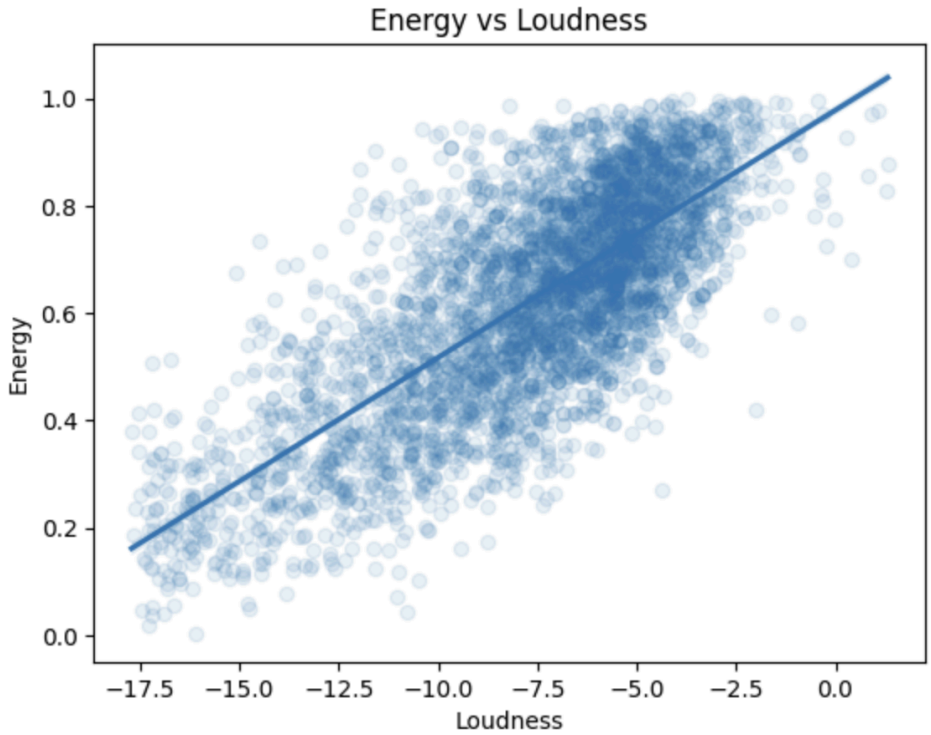
Then, we examined box plots representing multiple important numerical values that can be differentiated for High and Low-Ranking music:



Through this, we can see that the distributions for both high and low ranking songs are extremely similar for speechiness, valence, and tempo. However, it is important to acknowledge that the median of the high-ranked music is slightly higher in valence and slightly lower in speechiness, in comparison to the lower-ranked songs. In addition to this, lower-ranked songs have a slightly larger IQR for tempo, than higher-ranked songs. This means that although they have a similar range, tempo varies more in lower-ranked songs than in higher-ranked songs.

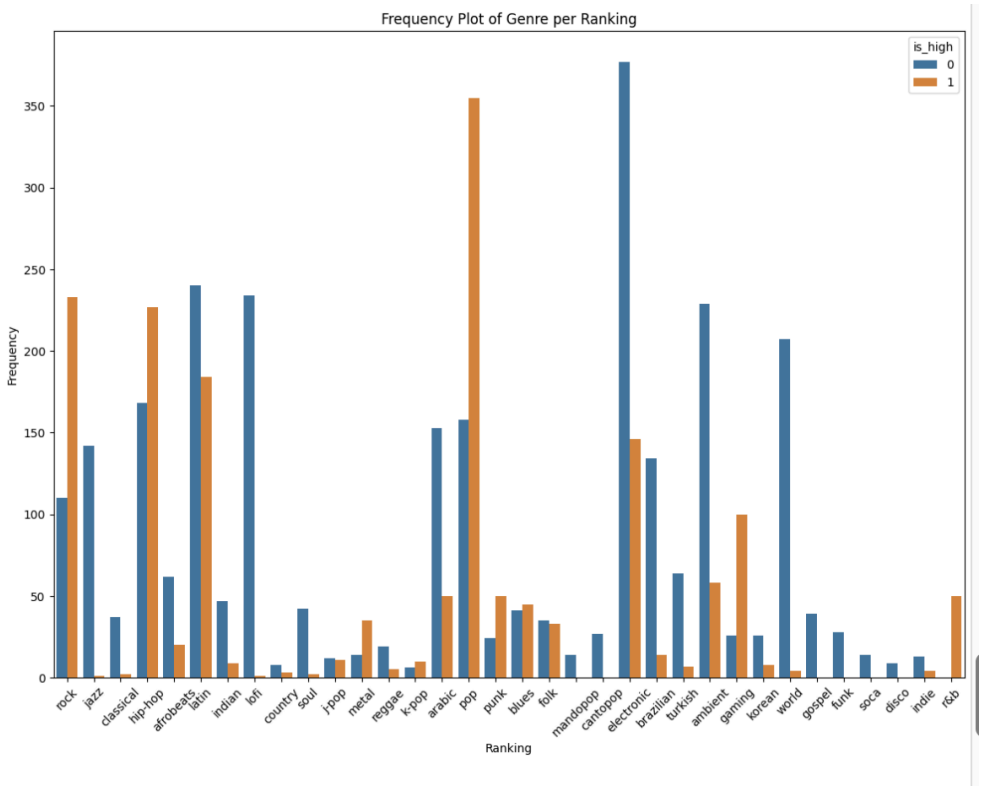
The loudness distribution displays the most difference between the ranked songs. The median for higher-ranked songs is larger than lower-ranked songs. This is also supported by the extreme amount of outliers that fall below the higher-ranked song distribution. This indicates that higher-ranked songs typically have a louder track in terms of decibels.

Next, we wanted to compare energy versus loudness:



As seen so far, we have discovered that higher-ranking songs typically have higher energy and loudness levels compared to lower-ranking songs. Looking at the plot, although the regression line shows an upward trend, it does not guarantee that Energy and Loudness go hand in hand. However, there is a lot of clustering around the linear regression line, indicating that it is common to see energy and loudness increasing with one another. It is also important to acknowledge that the cloud of clustering occurs near (-6.0,0.7).

Then, we compared Genres and Ranking of songs:



Since most of the data lies in higher frequency categories, we will not go over the less frequent categories in as much detail. As seen in this plot, some of the genres that lower-ranked songs dominate are jazz, latin, lofi, arabic, electronic, brazilian, ambient, and world. While higher-ranked songs dominate rock, hip-hop, and pop. This makes sense because a lot of the higher-ranked songs come from genres which are mainstream and very well-known genres, while lower-ranked songs come from more niche genres.

After creating the graphs, we made a linear regression model, by designating the track_popularity column as our target and all the numerical features as our feature. When we looked at the average R-2 score, we got 0.0599, which means approximately 6 percent of the features contribute to the overall popularity of the song. We also wanted to look at the correlation coefficient for each individual feature. We saw that features like loudness, key, and danceability had a high correlation with the popularity of song with coefficients like 0.72, 0.64, and 0.62, whereas features like acousticness and instrumentalness had way less correspondence with song popularity with values of -3.8 and -4.2.

We can analyze the results of our random forest tree. R2 measures how well the model explains the variation in our target variable. Our value was 0.23 which means about 23 percent, which we thought was strong as it gives us meaningful information about certain corresponding factors, like loudness and key. Mean absolute error (MAE) tells us on average how far off our model's prediction is from the actual score. The lower the number the better our model, and we got 15.74 which we believed was a good indicator on how our model predicted data. Then, we calculated mean squared error which is similar to MAE but it gives more weight to larger errors. We got 19.35 which reinforces that our model did not make too many major mistakes.

Overall, based on our results, we can conclude that both our linear regression and random tree models performed reliability and efficiently. There was a strong level of accuracy and minimal error, considering that our dataset had many attributes. We were able to collect a large amount of information on how various features affect the popularity of a song, which can help in solving the problem presented earlier in the paper.

We believe that our results were pretty similar to what we initially expected and gave us good insight into the makings of music popularity. While the model's predictive power was limited, the project showed us the importance of thorough data processing and its role in uncovering patterns and trends. To conclude, we gained a deeper understanding of the challenges in modeling music popularity and the potential of using data-driven approaches in the music entertainment industry.