Joseph Slater
Prof. Matthew Mcniell
05/18/2023
Data Tools Final Paper

Introduction

For my analysis I worked with the spotify and youtube dataset. The dataset contains a lot of information about songs ranging from artist and release type to valence and tempo. So there is a large mix of categorical data and continuous data. The dataset contents are meant for analyzing songs, specifically characteristics of songs such as key, valence, and tempo (and much more). It also includes view and stream counts on youtube and spotify for getting a gist of general popularity. What I intend to do with this dataset is to analyze its information and draw new conclusions using different techniques we learned in class.

Glossary can be found at the end of the paper

Exploratory Analysis

When analyzing the data, the first thing I made note of was which columns had categorical values and which had numeric values. I pinpointed which columns would have little to no value and removed them to simplify the dataset. I figured if I even needed to use that information I could always reintroduce them. After removing columns deemed useless I started doing a simple analysis of the data. First I wanted to see a count of the different release types[1]. By looking at the chart provided above, it became clear that albums are the most common type of release, followed by singles and the compilations. This information, while not useful in the present, was something I kept in the back of my mind in case it became an interesting talking point for future analysis.

The next thing I wanted to grasp was the correlation between values. So I created a correlation matrix[2].Upon first look, there is not much here to look into further, but that's what caught my eye the most. Shouldn't danceability and energy be further connected? How about speechiness and instrumentalness? Features you would expect to have a high correlation did not have any at all (while of course some like loudness and energy did). This sparked some ideas I decided to use later in analysis.

Finally I wanted to get a column of genres, however that is far harder to find than initially expected. I tried using multiple APIs, but their content was not up to snuff. So I thought back to our first class and remembered the mention of web scraping. It took a while to find a site which gave any sort of decent information, but eventually Wikipedia saved the day. However, the genres got for to specific so I ended up dropping the idea

Ethical Audit The data was collected from Kaggle, stored as a csv. How the creator of the csv found the data is not provided. I am under the assumption that the creator constructed the dataset

using ethical means. The data does not contain the identity of any individual. It mentions bands only for analytic purposes, but gives no information of the members. Statistics such as Spotify streams and YouTube views are only on a counting basis, so it is unknown if any given individual contributed to the tally.

The dataset's purpose was to be downloaded and utilized in a user's personal analysis, as it can be assumed from the publication of said dataset on a well known public database, Kaggle. My personal use is purely ethical and for data analysis, answering questions about music as a concept. It is unknown if the data of the bands who own these properties have given consent, but that would be on the head of the initial creator. However, all the information on a given song is either public domain or can be inferred from the song itself, so it must be assumed the creators are okay with such a datasets creation.

There is a near infinite amount of music, and with it artists, genres, and songs. Creating a totally inclusive dataset is impossible. There may be a leaning to certain genres, however any given artist contained in the dataset cannot have more then 10 songs. No algorithms I performed on this dataset look for a particular bias to support a claim I had previously. If the data reject's my claim, I make sure to admit so as to not compromise the ethicality of my analysis.

The dataset is stored on Kaggle. It is up to the creator and the site's cybersecurity to protect the data and prevent inaccurate information from being adjoined. It can be assumed a site such as Kaggle has put these safeguards in place to prevent malicious individuals from tampering with this data, allowing the data to maintain integrity. The public showing of this data keeps it highly available, and no confidential information is included in this dataset. The dataset will only be kept on my personal machine for as long as the analysis is being conducted, after which it will be promptly deleted.

The dataset makes sure to give the creators of these properties their due diligence, both with a link to their Spotify page as well as link to the song on YouTube. On the datasets posting there are clear descriptions of each column to avoid misinterpretation. I will be using the dataset in a responsible manner for my analysis.

Analysis 1 - Loudness and Energy Relationship

It can be inferred that loudness and energy share a relationship, as they would seem to indicate similar things. But I wanted to prove this relationship exists. First I created a scatter plot to visualize the data[3]. It had a positive correlation and a boomerang-like curve. This graph indicates there is a positive relation between the two, but I wanted to prove it further, so I did a Pearson test, and the results were very positive. The correlation was 0.7448445555528429 and the p value was 0.0. This indicates an extremely significant strong relationship. So next I decided to do linear regression, which resulted in a slope which could have been predicted[4].

Next I wanted to try to predict energy from loudness. This model gave me a R-squared of 0.554499773890772, a MSE of 0.02023667796545928, and a MAE of 0.113605613315751. So the model offers a decent explanation of the variation while also offering accurate predictions. The model I decided to plot the line of regression of actual vs expected energy[5]. While you could predict energy from this, I wanted to improve it, so I removed outliers. I got rid of values more than 25% away from the median (so I kept from 25% - 75%)[6]. The Pearson test gave values of 0.6752643856032233 for correlation and 0.0 for p value. So the correlation actually got slightly weaker here. Regardless, I pressed on. I made the linear regression line[7]. Which looking at the data was to be expected. I ran another linear regression and the results were r-squared of 0.4652220406213581, MSE of 0.01760331554382998, and MAE of 0.10677881275805037.

So what did this all mean? Well, the outlier variant performed better according to r-squared. However, the non-outlier version performed better when it came to MSE and MAE. Although it should be noted there was a significant drop in r-squared, while only a minimal drop in MSE and MAE. So for future use, I would use the version with outlier as opposed to the one without it.

Analysis 2 - KNN with Tempo, Valence, Danceability, Key

Thinking about music, I figured there would be some relationship here where songs of a certain mood (valence), tempo and danceability would link to a particular key. So I decided to attempt to make a KNN to find songs with similar key. So, I first changed the Key values from a number to the value the number represents. I then ran a ANOVA test. Accousticness and danceability performed well so I was only going to use those two for my analysis, but I wanted to dig deeper. I ended up making boxplots for all numerical columns to compare against the Key column[8].

After analyzing both the boxplots and the ANOVA results, I decided to work with valence and accousticness. I created scatter plots for the valence accousticness relationship[9] against all the categories of key. After being unhappy with the results, I decided to keep looking. So I tried danceability and valence[10]. I was far happier (but not thrilled) with the results of these scatter plots. I still tried danceability and accousticness[11] just to confirm danceability and valence was what I wanted. After looking at all 3, it still looked the best so I ran with it.

I changed the keys back into their numeric values, and then made a KNN with n neighbors of 3. I tried 4 and 5 as well, but none seemed to perform well. It peaked at an accuracy of 0.1541988416988417. So I decided to try every option. First tried just valence and just danceability. Both performed poorly, with values of 0.10762548262548262 and 0.10762548262548262 respectively. Still unimpressed, I tried accousticness, and it only scored marginally better at 0.12210424710424711. So in the end, my conclusion was that valence, danceability, and accousticness could not perform KNN on key well.

Analysis 3 - Linear Regression of Popularity

In the initial preprocessing, I separated the views, stream, likes, and comments columns to their own dataframe. I figured they are all connected and I can work with them later, and so I did. I thought that these four columns would be plausible to make predictions on one another as generally if a song is popular on one platform in one way, it can be popular on another platform or the same platform in another way. So I decided to use views, which refers to youtube views as my initial basis as that was the thread that connected them together.

First I looked at the data and noticed massive outliers. So I tried three different ways of shrinking the datapool and decided on the range of 25% to 50%, which had a fair range on all values. So I went ahead and looked at the relationship between views and popularity[12]. What I saw did not impress me, so I did a Pearson test. I got a correlation of 0.022757792322625578 and a p value of 0.4867957470144937. Essentially, we had nothing here. Nonetheless, I decided to make the regression model[13], and to the shock of no one, it performed miserably.

Next I analyzed likes vs views. I figured if a youtube video has a bunch of likes, it has to be popular. So I created another scatter plot of the data[14]. This one also didn't look promising, but I still did a Pearson test. These results were much better, with a correlation of 0.30797714844927443 and a p value of 5.156178274543634e-22. These values indicate there is a small positive relationship that is a line between significant and insignificant. So I made the linear regression model[15] and it also performed very poorly with a r-squared of 0.10672523662455047.

I repeated the process with comments[16][17]. Once again it failed. Finally I tried a new base. I thought maybe comments and likes would have a better relationship. It did not[18]. I eventually came up with the idea of scaling down all the values by min max[19], PCA[20], log1p[21], and Box Cox[22]. None of the scatter plots looks remotely good visually, so I stopped trying Pearson and linear regression.

While I was disappointed I didn't find a relationship, the lack of a relationship is interesting in of itself. There is no real connection between Spotify and Youtube performance. There is no link between popular songs and comments or likes. And there's no link between likes and comments. Essentially it's random how well you perform in each category. While yes, if you do well on spotify odds are you will do better on YouTube. And if you have hundreds of

thousands of likes you're probably going to have more comments then a video with 50,000 likes. But there is no overarching pattern between them all.

Analysis 4 - Linear Regression between Accousticness and ‘Intensity’

Thinking about this logically, there should be an inverse relationship between acoustic songs and ‘intense’ songs, so I decided to analyze this. First I wanted to construct the intense column. I decided merging the energy and loudness column would do the trick. First I summed the columns before diving by 2 (essentially finding the average), then plotten the relationship to visualize these changes[23]. Looking at the Pearson test, with the values of -0.6609326152589744 and 0.0 for correlation and p value respectively, it is clear there is a significant negative relationship here.

Next I decided to apply linear regression to this relationship[24]. The values I got were r-squared is 0.44954223178301345, MSE is 0.01180679792857349, and MAE is 0.08450059128511306. Essentially, nearly half the variance of intensity could be explained by accousticness. So I proceeded to plot the actual vs expected intensity[25]. I suppose the slope likes being positive, because I’m not exactly sure why the graph flipped upside down, but the data was the same so it made no difference.

After getting differing results previously from removing outliers earlier, I decided to try it again. So I made the same split and plotted the results[26]. After seeing the scatter plots have a different shape, I decided to run the linear regression again[27]. With a r-squared of 0.3560326249519006, MSE of 0.04561923413989519, and a MAE of 0.17133009811405317, the outlier model performed worse than the non-outlier model[28]. So while neither model was amazing, the first model is good enough to get a reasonable prediction.

Analysis 5 - Clustering Danceability and Valence

This test was done just purely from a curiosity perspective. In my experience, music that is uplifting is good to dance too. But was that really true? I wanted to see what I can learn from clustering these features. So first I created my new dataframe from the danceability and valence columns., then plotted these points to look for a relationship[29].

After analyzing the scatter plot, I decided that clustering could find some interesting results. So I conducted a clustering with 4 centroids[30]. Testing was also done with centroid values of 3 and 5, but 4 seemed to give the most interesting results, so those were the ones I left. Before getting into analysis, I had an idea. I wanted to try sampling a random group of 1000 entries to see if that changed the results. After doing so, I also made a cluster of 4 centroids for that and it results in a very similar plot[31]. While this is expected, it made the plot clearer and more readable (to me at least).

So, what conclusions did I draw? Well, I decided to analyze the groups of clusters by the mean of their valence and danceability. What I saw surprised me. The best music for dancing was the happy mood music, which is to be expected. However, the second best genre wasn't the second happiest song, it was actually the second saddest. My logic was that ballad type songs which are good for dancing are generally on the sadder sounding side, even if the lyrics aren't necessarily and they are not "down in the dumps". They hit a perfect threshold of being sad but not too sad, and that made them perfect for dancing. The next category was the second happiness, and then came the precipice drop to the fourth of pure sadness. Here is a bar chart showing off the results[32].

Analysis 6 - Does Album Type have an Effect on Streaming Numbers?

When I think of music, I generally think of the big singles artists release before an album to hype them up. Usually what ends up being one of the most popular songs on an album is released as a single. Would singles on average have higher streaming numbers then songs on albums? I wanted to see if that was true. First I made a bar chart of the album types to visualize the data[33]. It gave me an idea of what I was looking at. While yes, there are far more album songs, that was to be expected as an album is a collection of songs while a single is…well…a single song.

So I made a new dataframe for simplicity sake and then did an aggregation of grouping album type and looking for the mean of stream. These were the results[34]. I found this very interesting, but I wanted to verify it so I performed an ANOVA test. The test gave results of a f statistic value of 83.95612021906973 and a p value of 4.943599334792968e-37. This means these results of this analysis were true.

What do these results mean? It means streaming numbers are actually *better* in albums than in singles. My assumption was incorrect, and I have a theory to back it with. I believe some people listen to their favorite songs in the album and let it play out, increasing the mean for albums. However, for singles some flop, and that can tank the entire performance of the single album type. To be honest there was no explanation for what a compilation was, and I liked having 3 bars so I left it despite ignoring it.

Conclusion

So what can be learned from this dataset? Let's look at the analysis 1 by 1. In analysis 1, it became clear the loudness and energy contain a strong relationship, strong enough to create a linear regression model. It's fair to assume in general if a song is loud it's also bringing the energy. This can possibly be due to powerful vocals and loud instruments being key features in both data groupings.

The second analysis showed that the data of danceability, valence, and accousticness were insufficient at predicting the key of a song although all of these categories are often associated with specific keys. Even though these categories are all associated with specific types and keys of music, there is still a significant range they do scope too, and this range makes something like KNN impossible in this scenario.

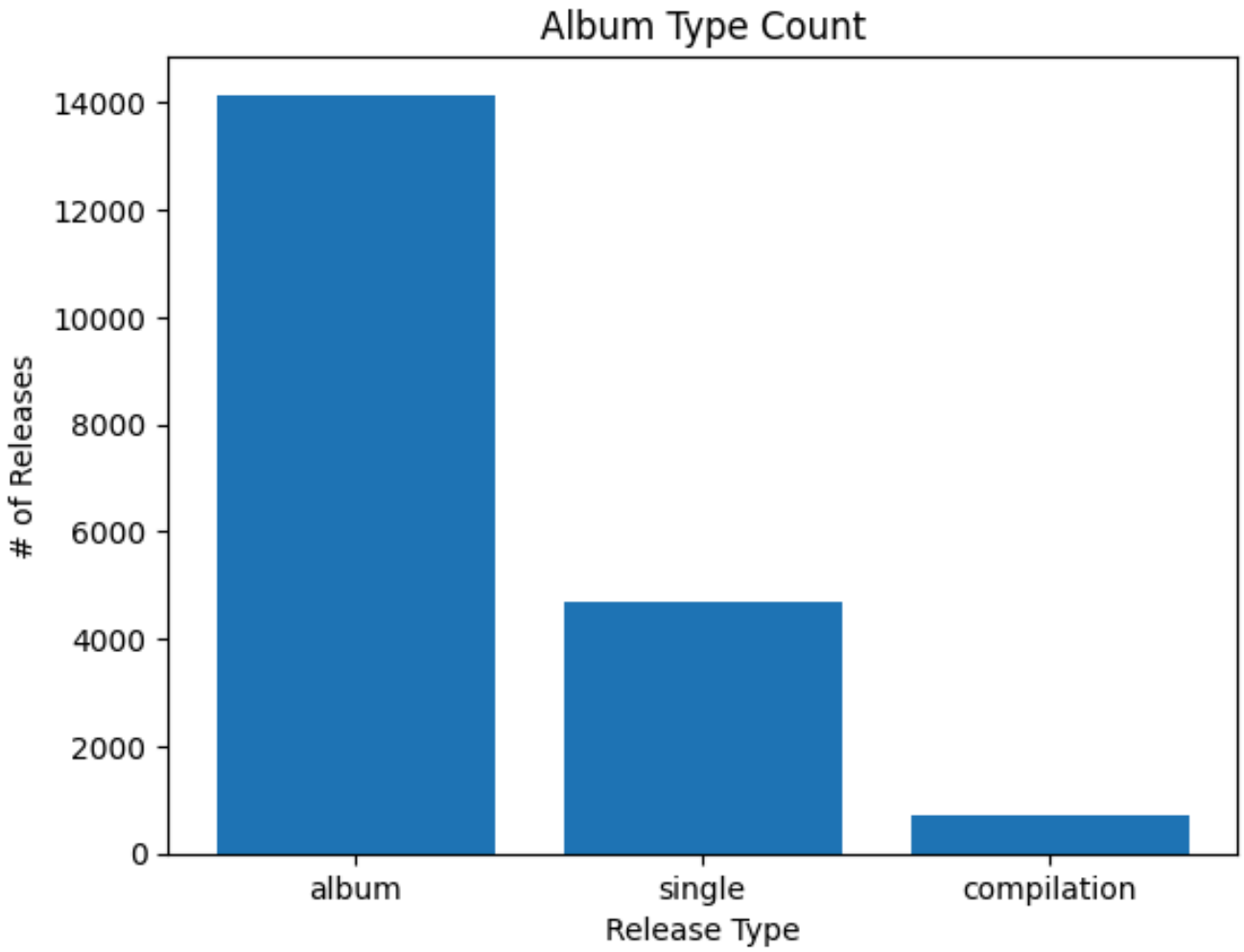
Third it became clear that the popularity of music is not shared both on different platforms such as Spotify and Youtube as well as in different measurements on the same platform such as likes and comments. Someone who uses Spotify won't be likely to also listen to music on YouTube, so it makes sense that their usage counts differ. And not all users use a platform the same. Some love liking, some never do, and the same could be said for comments which explains the lack of cohesion.

In the fourth analysis the evidence pointed to the fact that accousticness and intensity share a significant negative relationship strong enough for a competent linear regression model. This is likely due to acoustic instruments having gentle sound paired with soft vocals while intensity comes from intense instruments like electric guitar and ear piercing vocals.

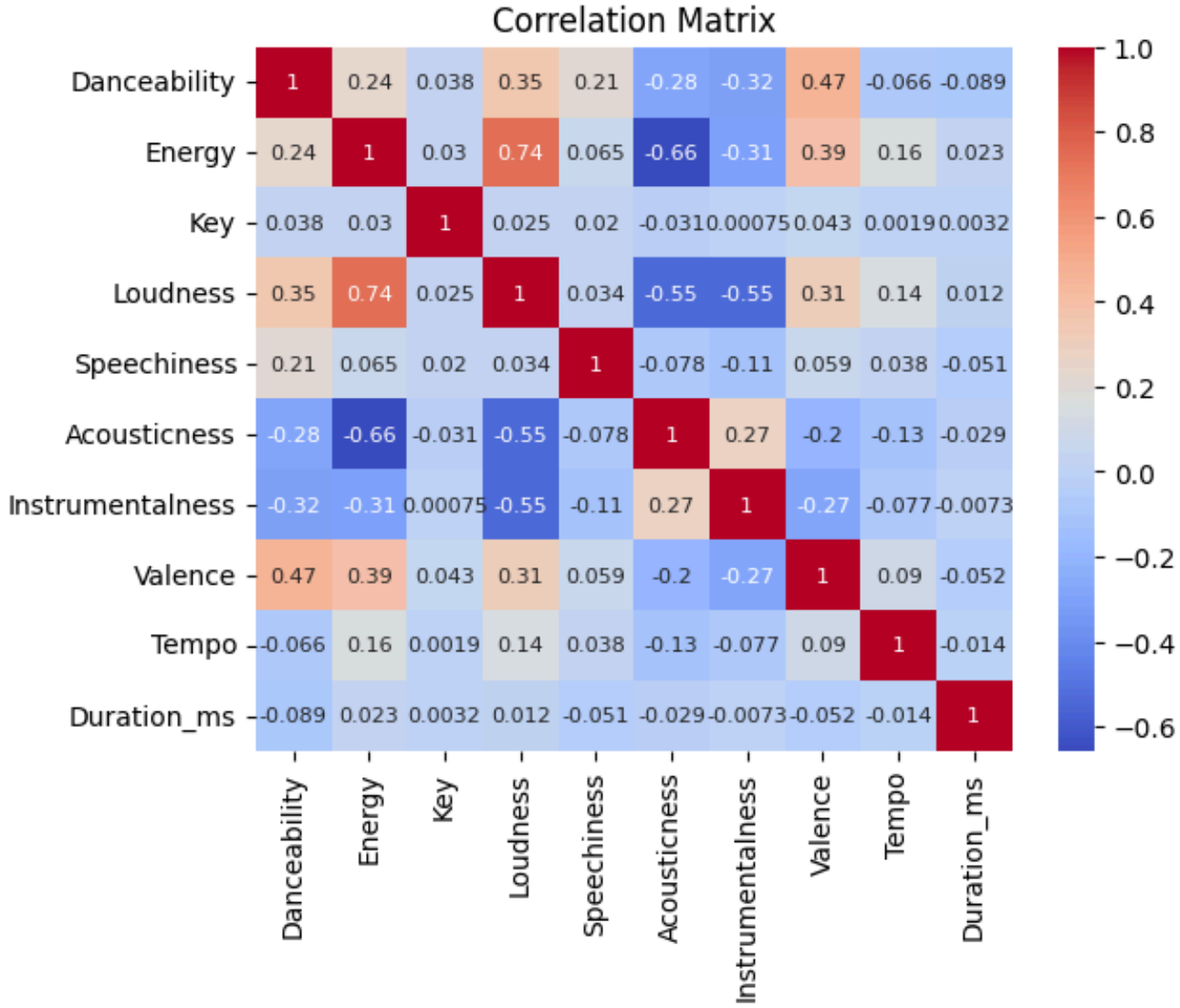
During the fifth analysis we learned that valence and danceability do not have a linear relationship. There is a threshold of sadness that beats mediocre happiness in regards to danceability. Finally in the last analysis it became clear that despite my initial thought, songs in an album are more popular on Spotify (and likely all music streaming platforms) than singles in regard to quantity streamed. This could be due to the fact the albums are bundled, so if you listen to 1 song and it plays out, users are likely to listen to the next and so on. If you want to listen to a single you need to actively seek it out

Glossary

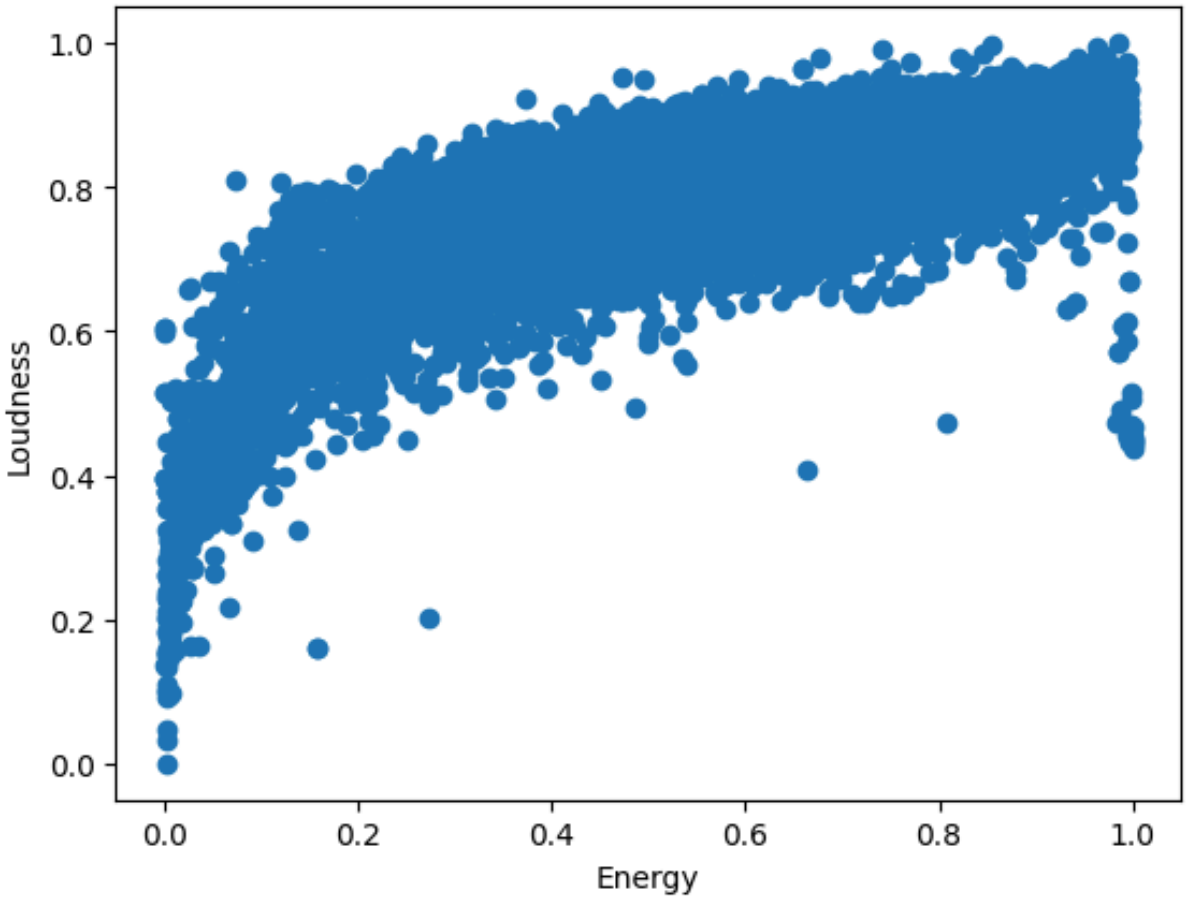
1.


Album Type Count
14000
12000
10000
8000
6000
4000
2000
0
# of Releases
album
single
compilation
Release Type

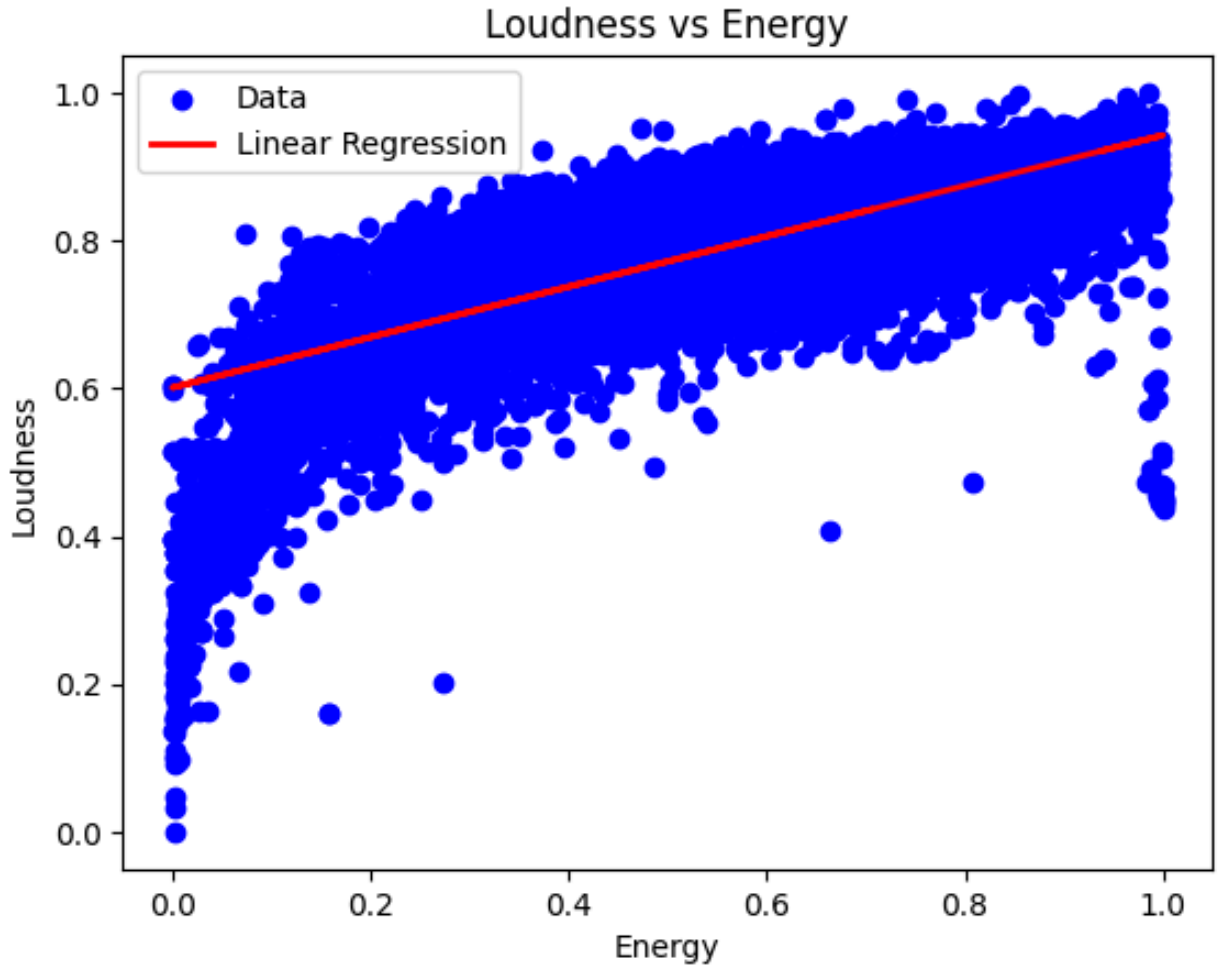
2.


Correlation Matrix
Danceability 1 0.24 0.038 0.35 0.21 -0.28 -0.32 0.47 -0.066 -0.089
Energy 0.24 1 0.03 0.74 0.065 -0.66 -0.31 0.39 0.16 0.023
Key 0.038 0.03 1 0.025 0.02 -0.031 0.00075 0.043 0.0019 0.0032
Loudness 0.35 0.74 0.025 1 0.034 -0.55 -0.55 0.31 0.14 0.012
Speechiness 0.21 0.065 0.02 0.034 1 -0.078 -0.11 0.059 0.038 -0.051
Acousticness -0.28 -0.66 -0.031 -0.55 -0.078 1 0.27 -0.2 -0.13 -0.029
Instrumentalness -0.32 -0.31 0.00075 -0.55 -0.11 0.27 1 -0.27 -0.077 -0.0073
Valence 0.47 0.39 0.043 0.31 0.059 -0.2 -0.27 1 0.09 -0.052
Tempo -0.066 0.16 0.0019 0.14 0.038 -0.13 -0.077 0.09 1 -0.014
Duration_ms -0.089 0.023 0.0032 0.012 -0.051 -0.029 -0.0073 -0.052 -0.014 1
Danceability
Energy
Key
Loudness
Speechiness
Acousticness
Instrumentalness
Valence
Tempo
Duration_ms
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4
−0.6

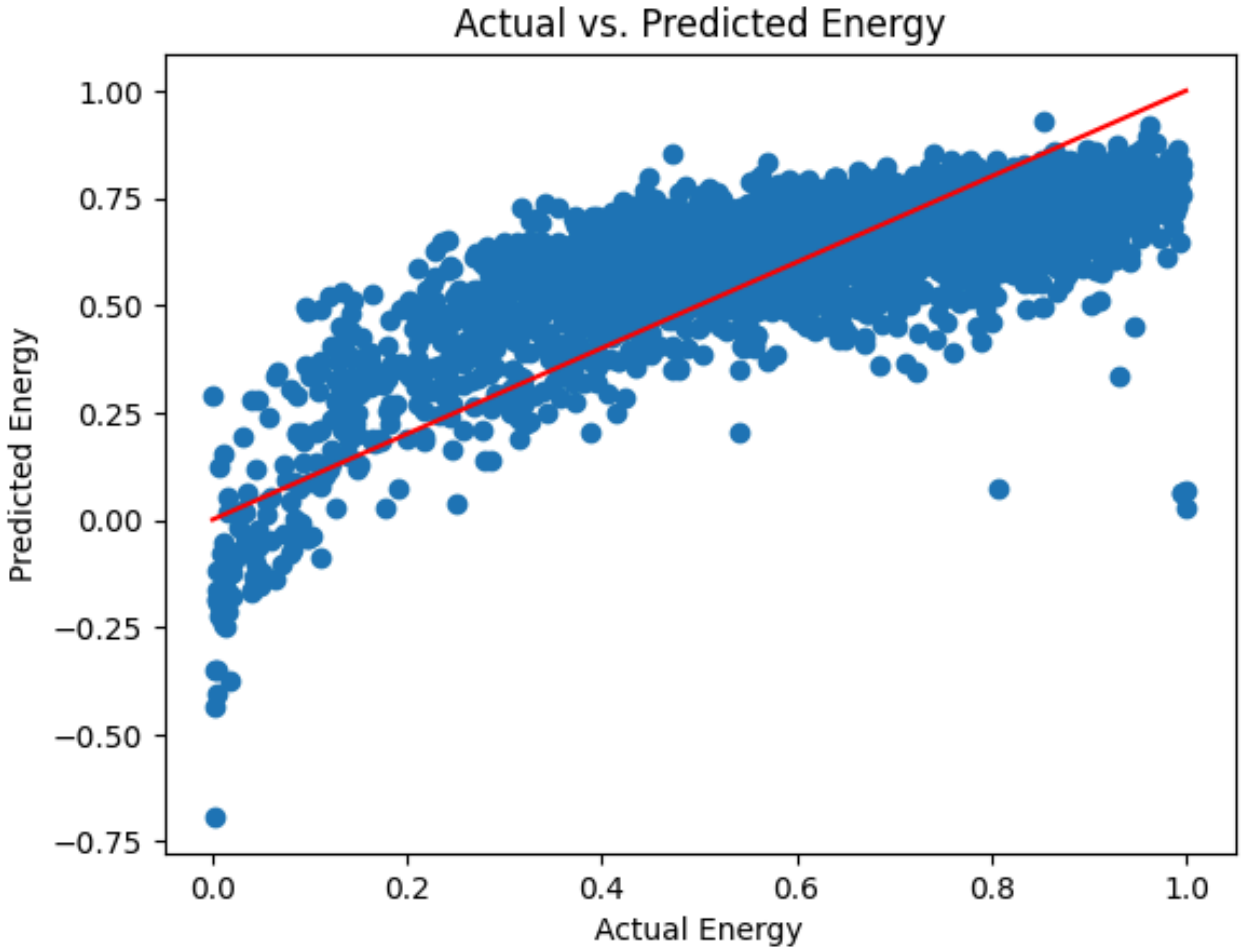
3.



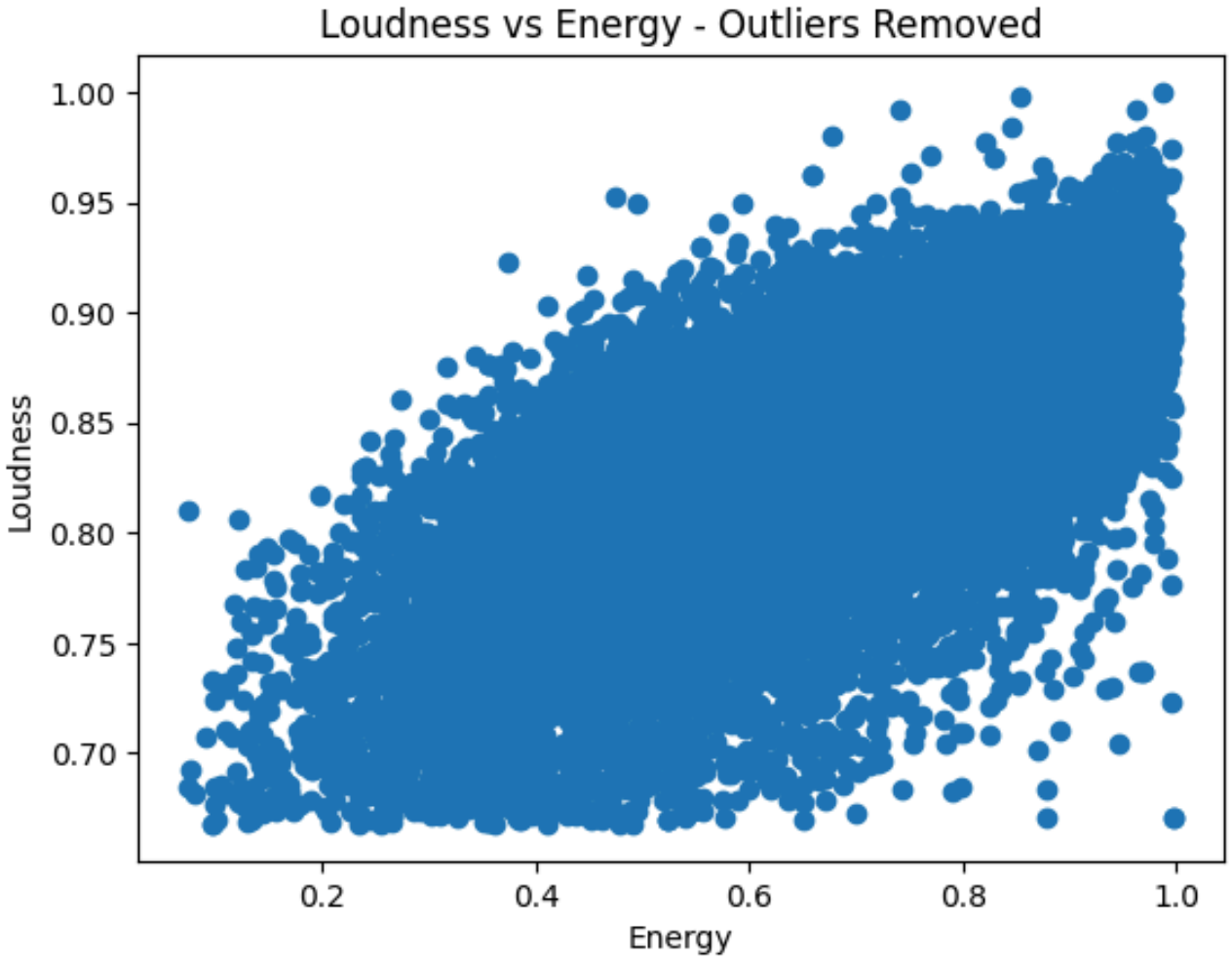
4.


Loudness vs Energy
Data
Linear Regression
Loudness
Energy
0.0
0.2
0.4
0.6
0.8
1.0

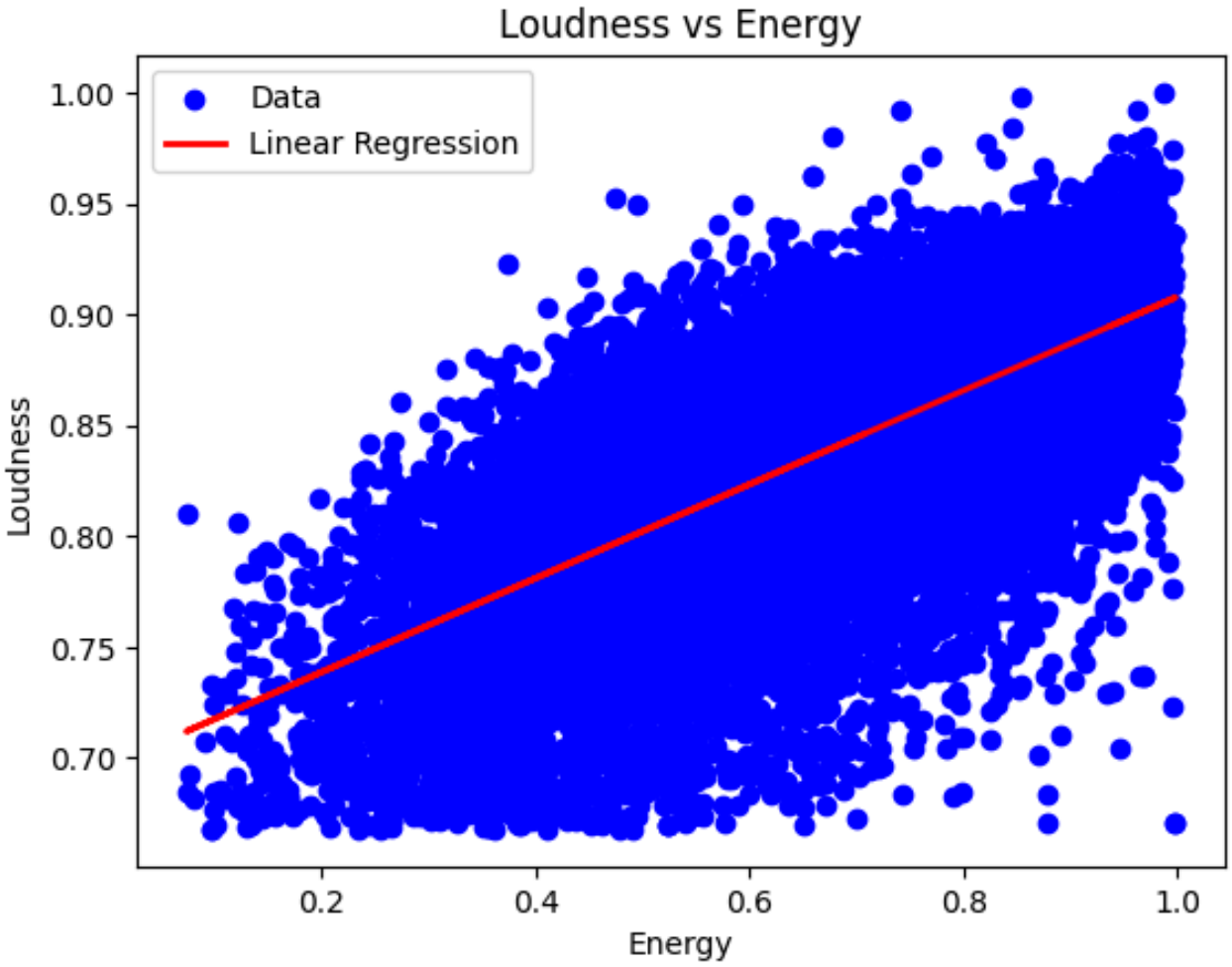
5.


Actual vs. Predicted Energy
Predicted Energy
Actual Energy
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
0.0
0.2
0.4
0.6
0.8
1.0

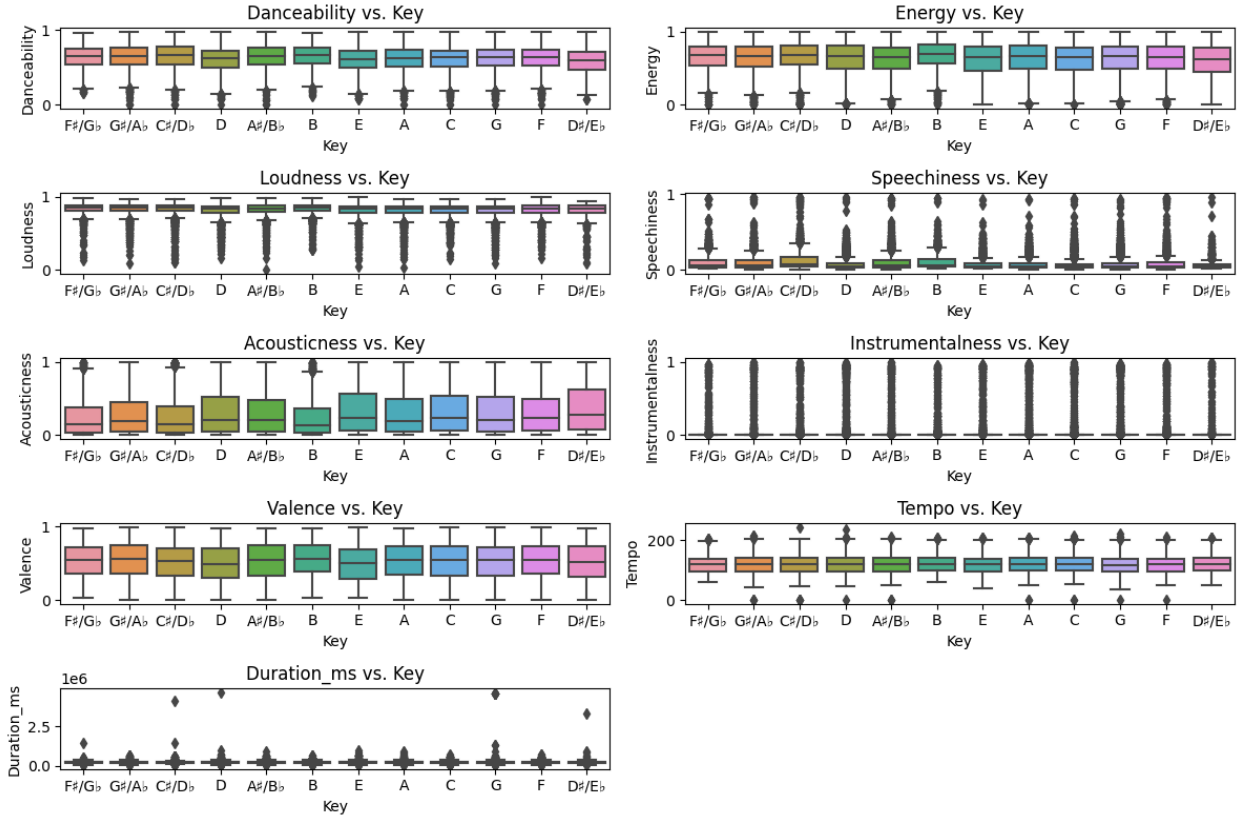
6.


Loudness vs Energy - Outliers Removed
Loudness
1.00
0.95
0.90
0.85
0.80
0.75
0.70
0.2
0.4
0.6
0.8
1.0
Energy

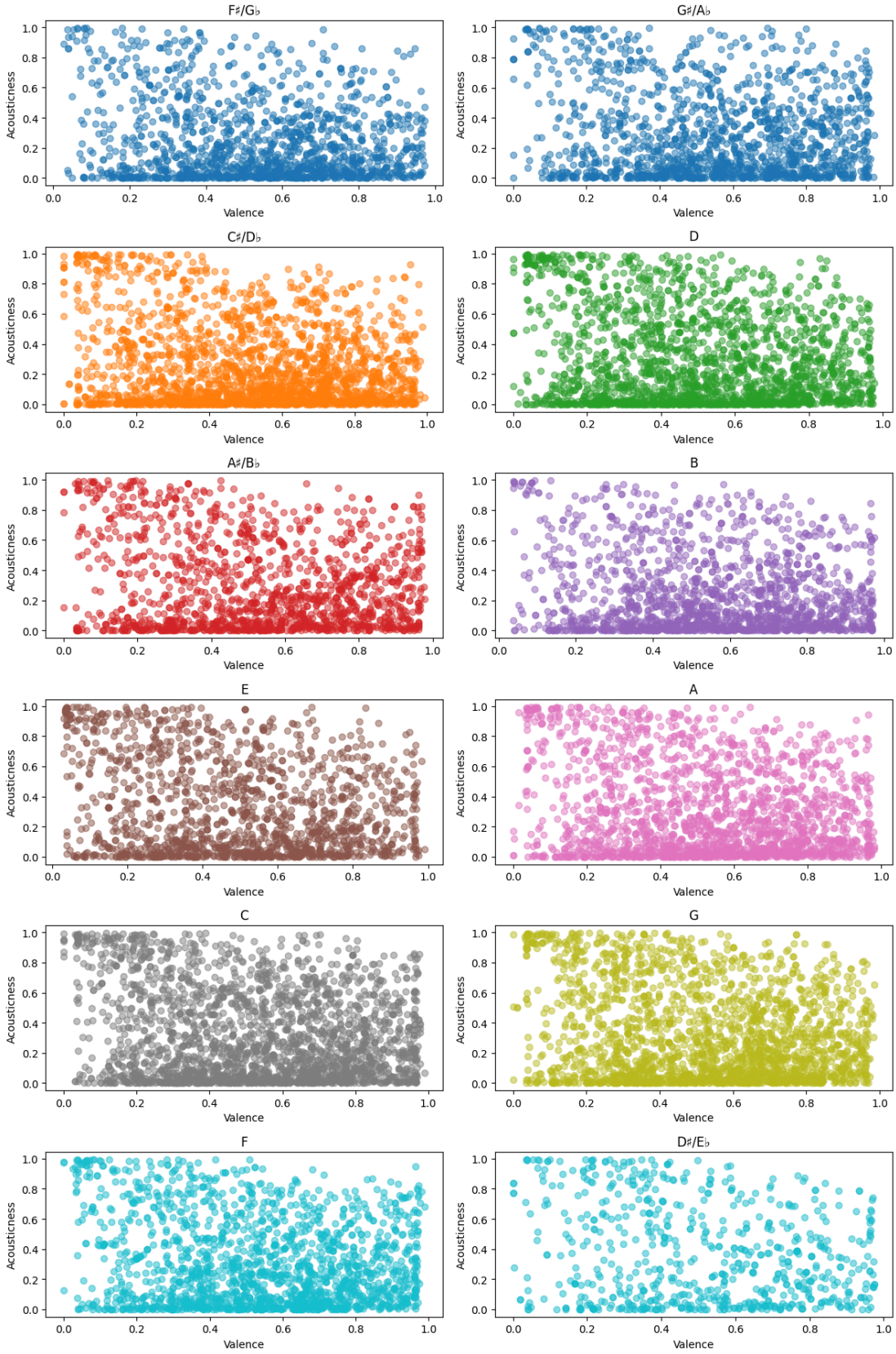
7.

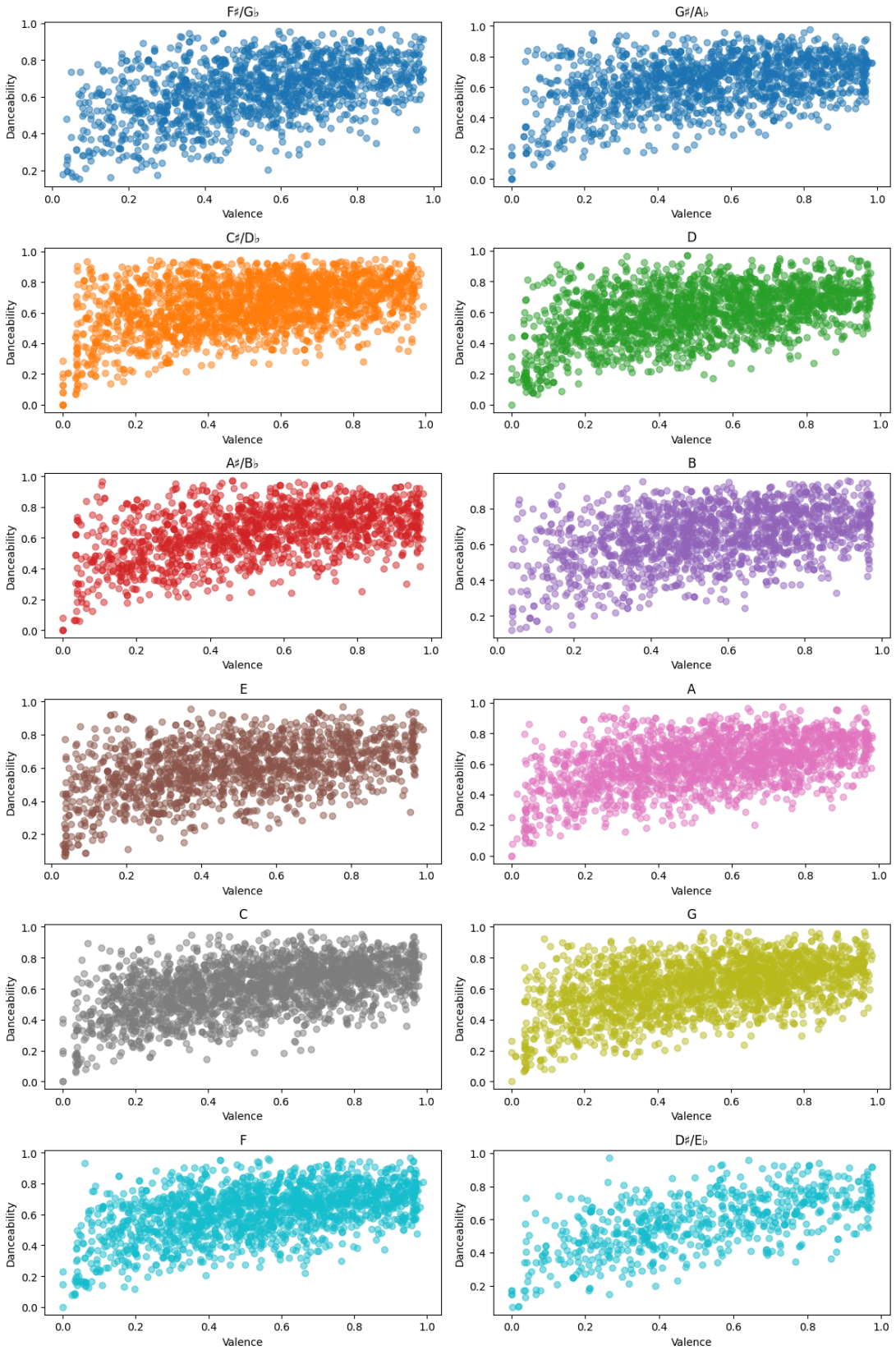

Loudness vs Energy
Data
Linear Regression
1.00
0.95
0.90
0.85
0.80
0.75
0.70
Loudness
0.2
0.4
0.6
0.8
1.0
Energy

8.

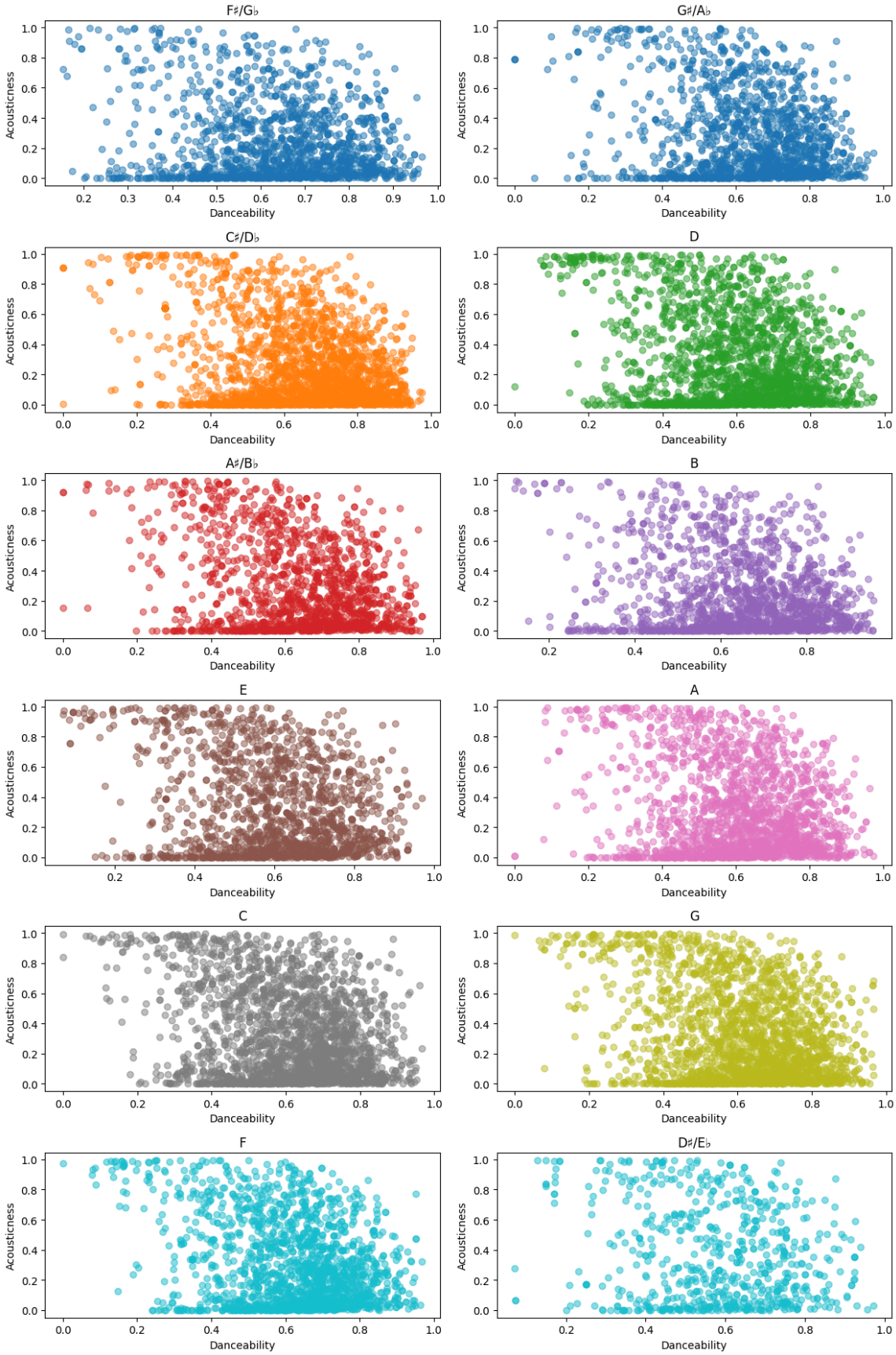

Danceability vs. Key
Danceability
Energy vs. Key
Energy
Loudness vs. Key
Loudness
Speechiness vs. Key
Speechiness
Acousticness vs. Key
Acousticness
Instrumentalness vs. Key
Instrumentalness
Valence vs. Key
Valence
Tempo vs. Key
Tempo
Duration_ms vs. Key
Duration_ms
1e6
Key
F♯/G♭ G♯/A♭ C♯/D♭ D A♯/B♭ B E A C G F D♯/E♭

9.



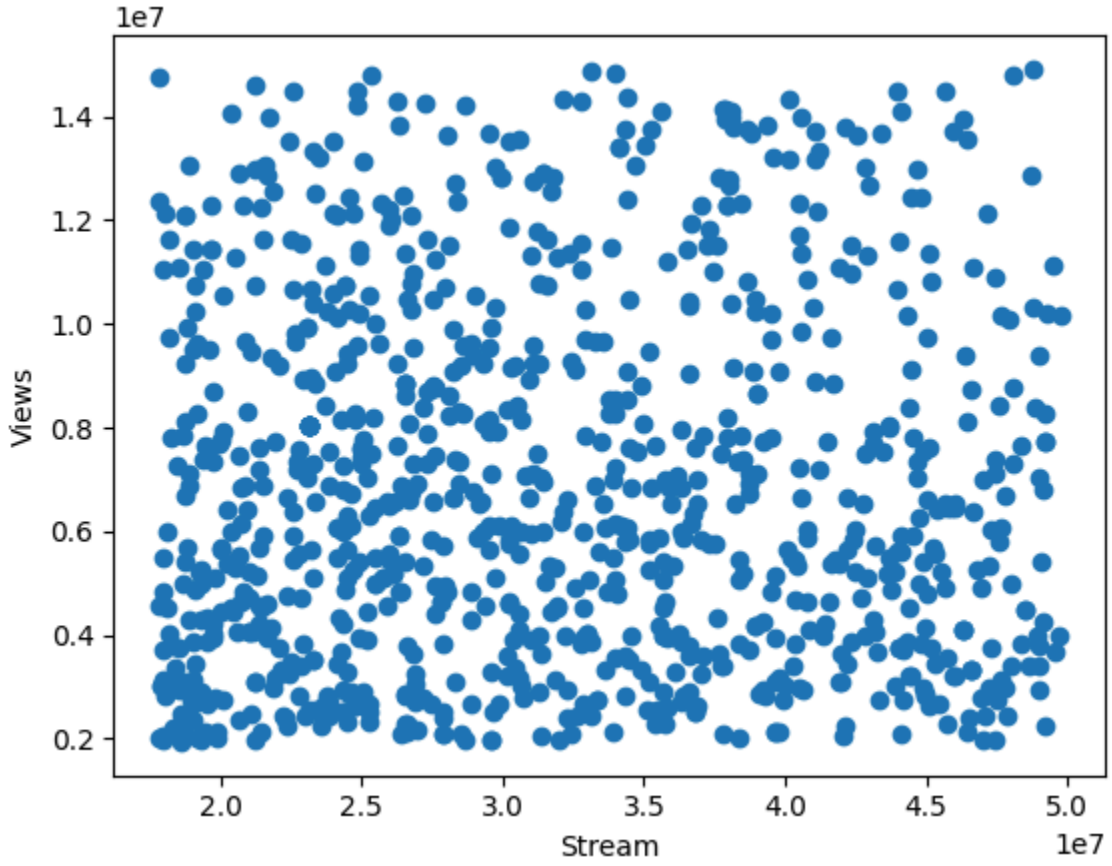


10.

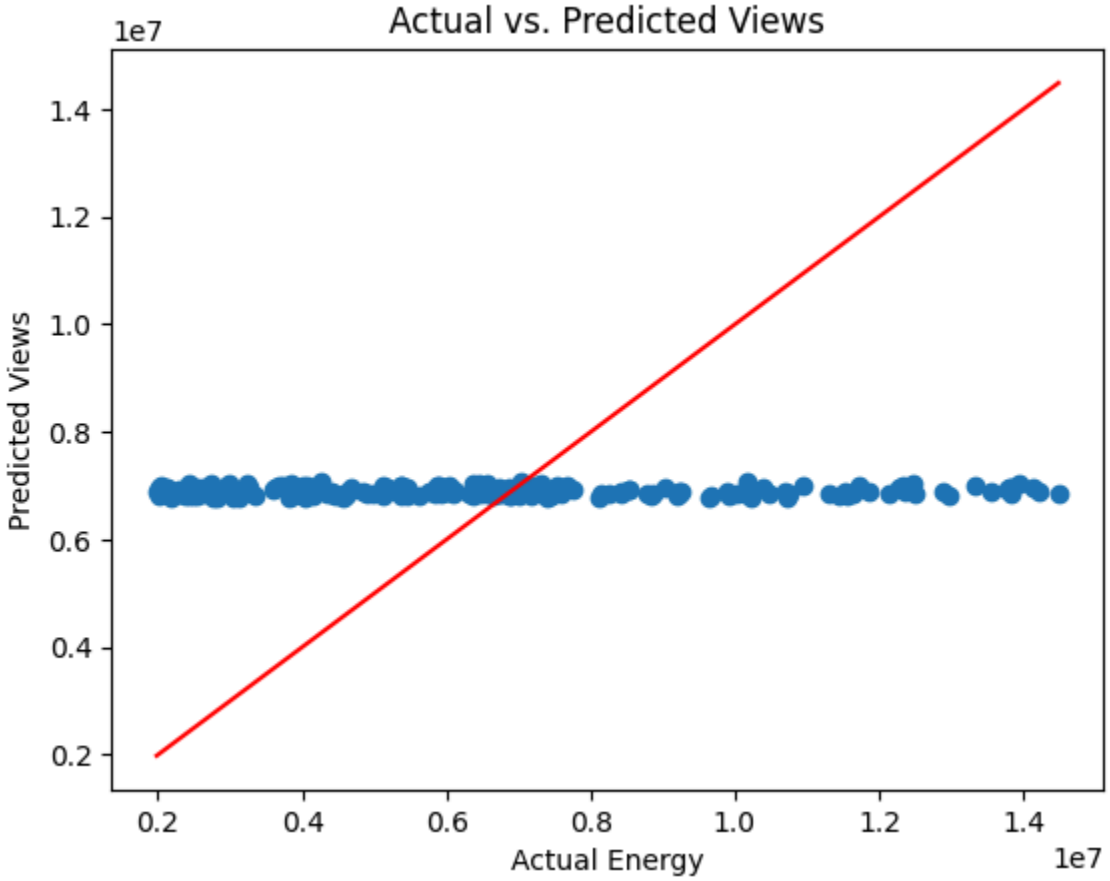


11.

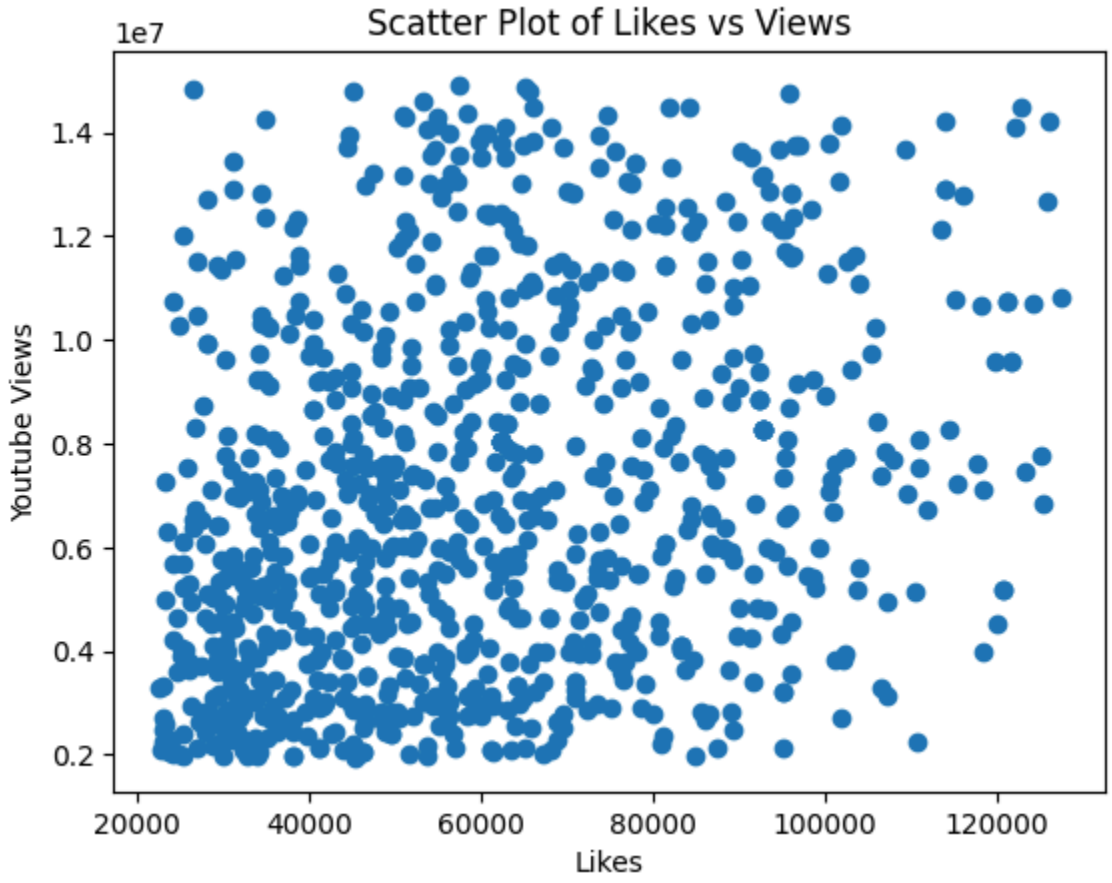
12.



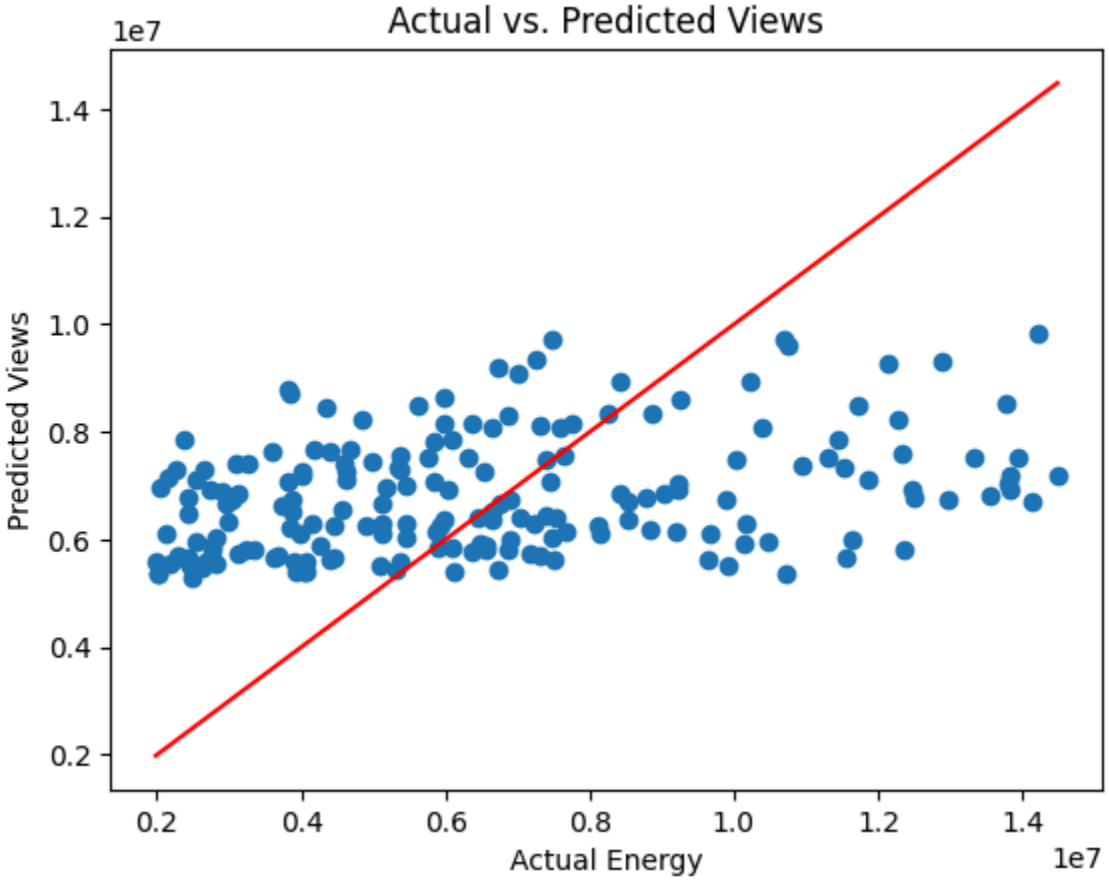
13.


Actual vs. Predicted Views
1e7
1.4
1.2
1.0
0.8
0.6
0.4
0.2
Predicted Views
0.2
0.4
0.6
0.8
1.0
1.2
1.4
Actual Energy
1e7

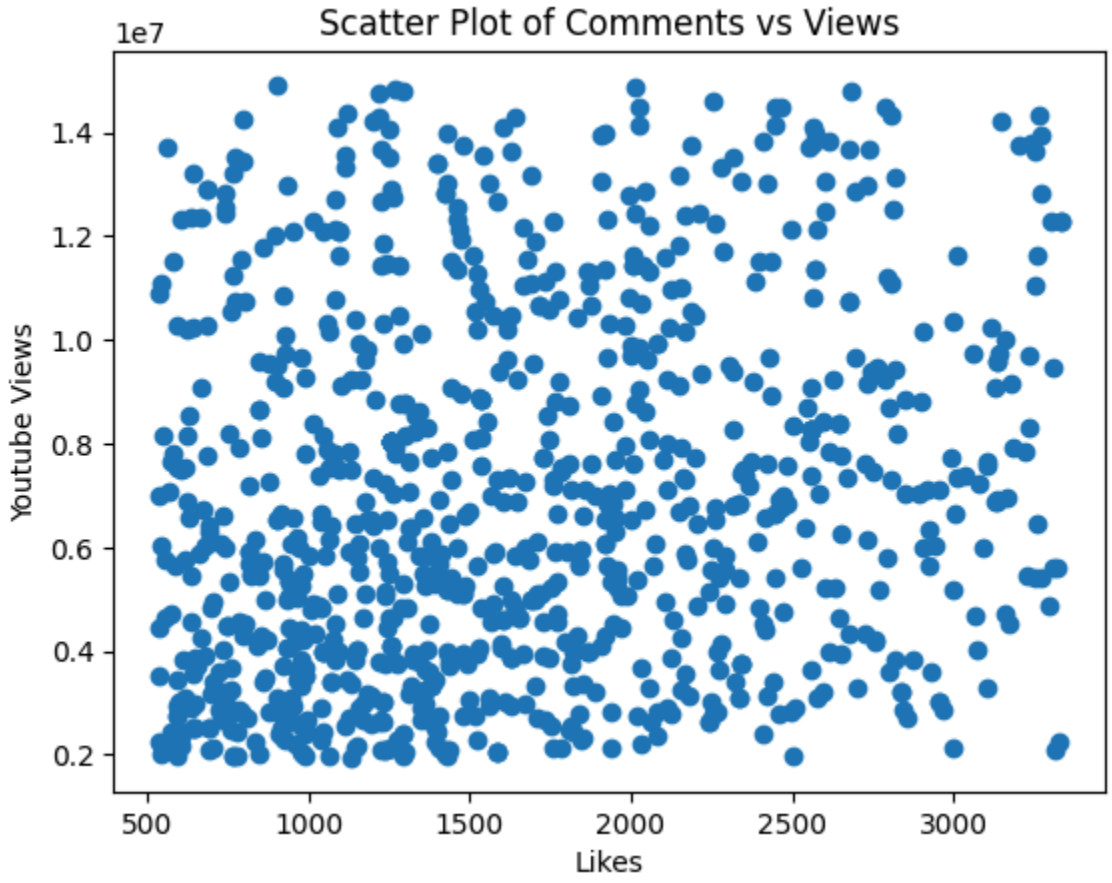
14.


Scatter Plot of Likes vs Views
1e7
1.4
1.2
1.0
0.8
0.6
0.4
0.2
Youtube Views
20000
40000
60000
80000
100000
120000
Likes

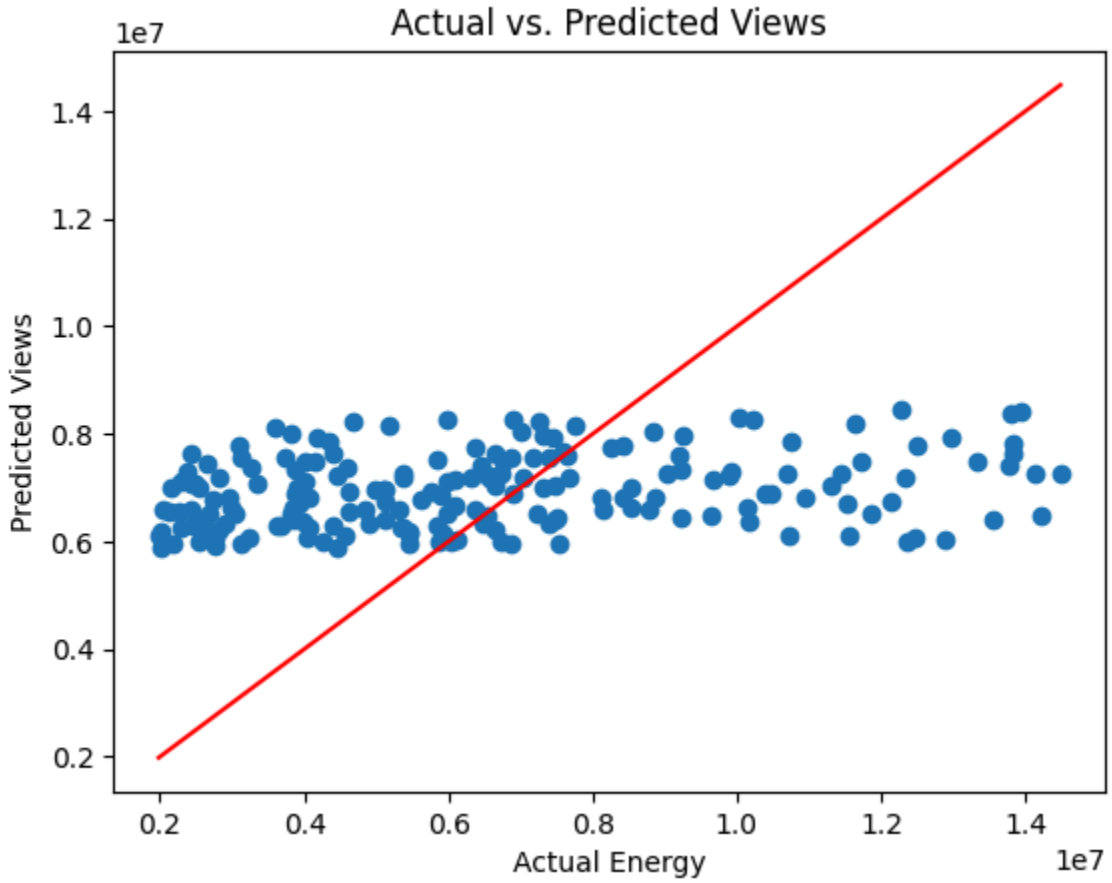
15.


Actual vs. Predicted Views
1e7
1.4
1.2
1.0
0.8
0.6
0.4
0.2
Predicted Views
0.2
0.4
0.6
0.8
1.0
1.2
1.4
Actual Energy
1e7

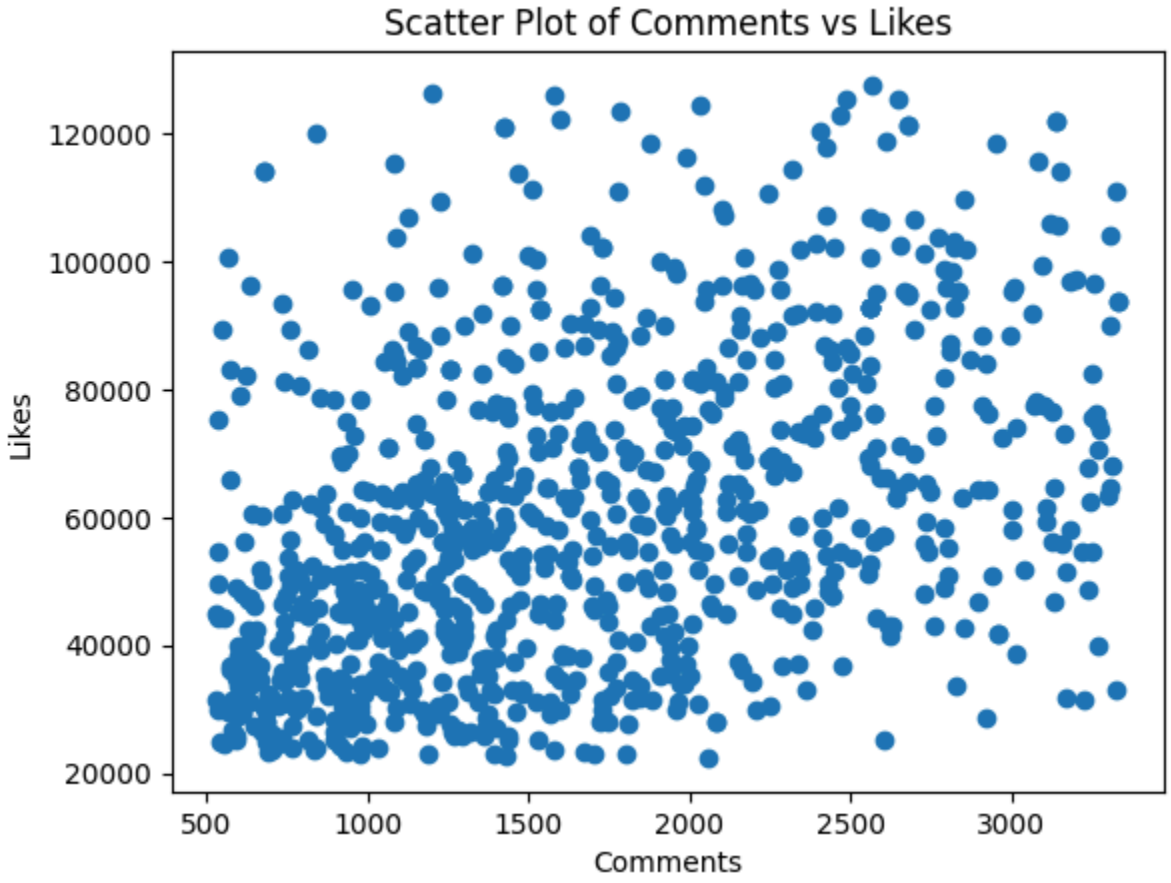
16.


Scatter Plot of Comments vs Views
1e7
1.4
1.2
1.0
0.8
0.6
0.4
0.2
Youtube Views
500
1000
1500
2000
2500
3000
Likes

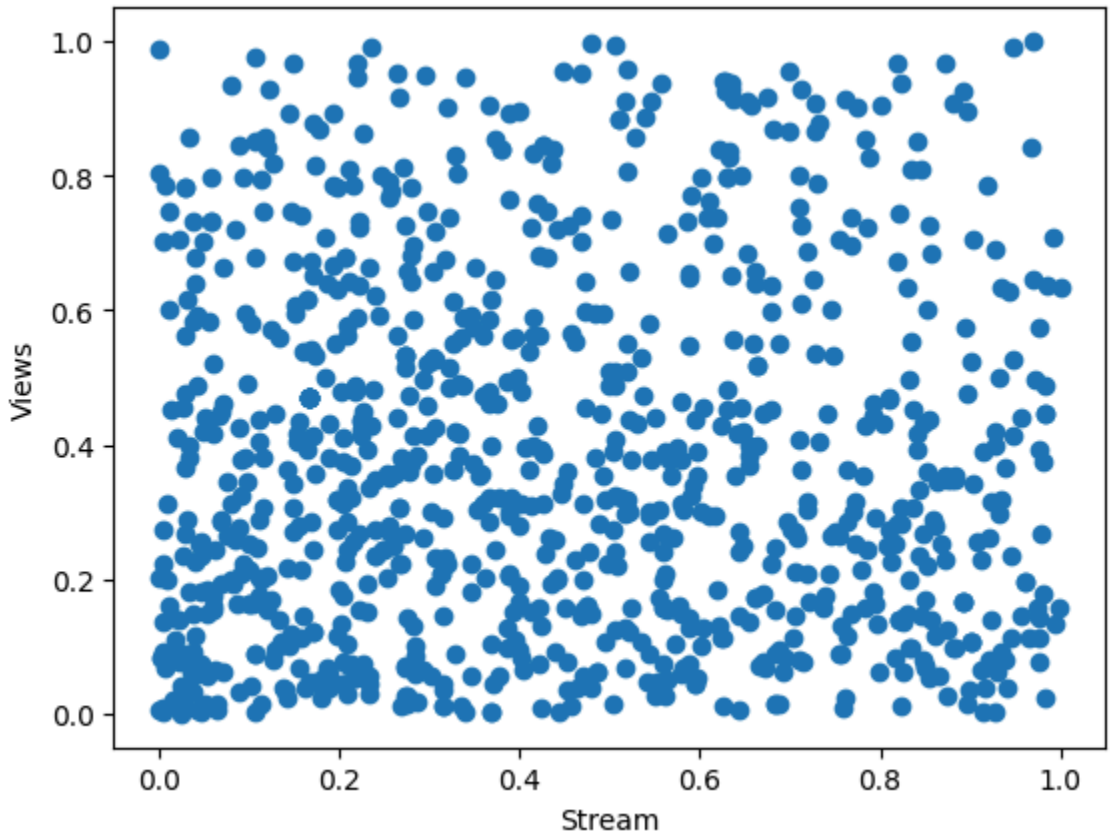
17.


Actual vs. Predicted Views
1e7
1.4
1.2
1.0
0.8
0.6
0.4
0.2
Predicted Views
0.2
0.4
0.6
0.8
1.0
1.2
1.4
Actual Energy
1e7

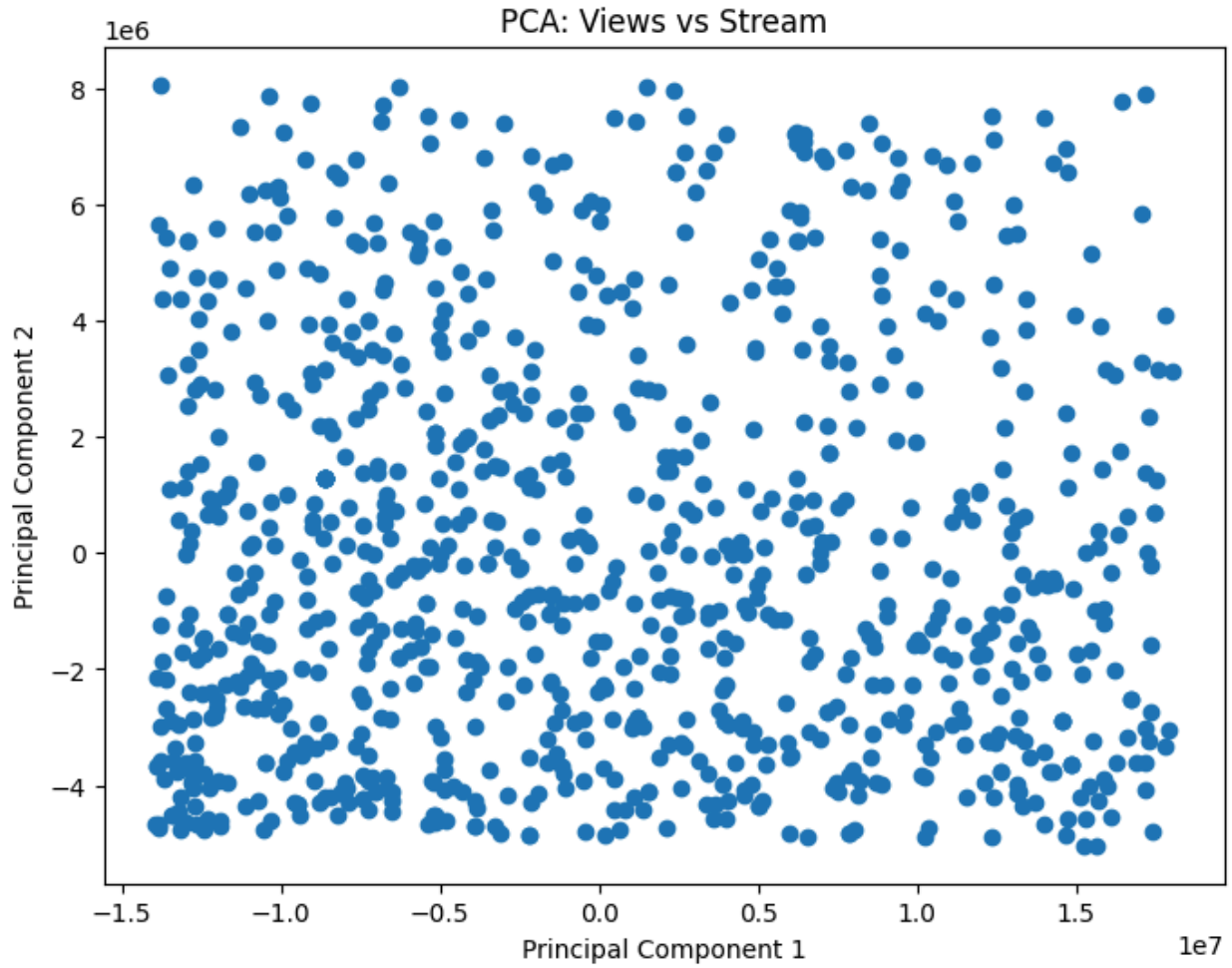
18.


Scatter Plot of Comments vs Likes
Likes
120000
100000
80000
60000
40000
20000
500
1000
1500
2000
2500
3000
Comments

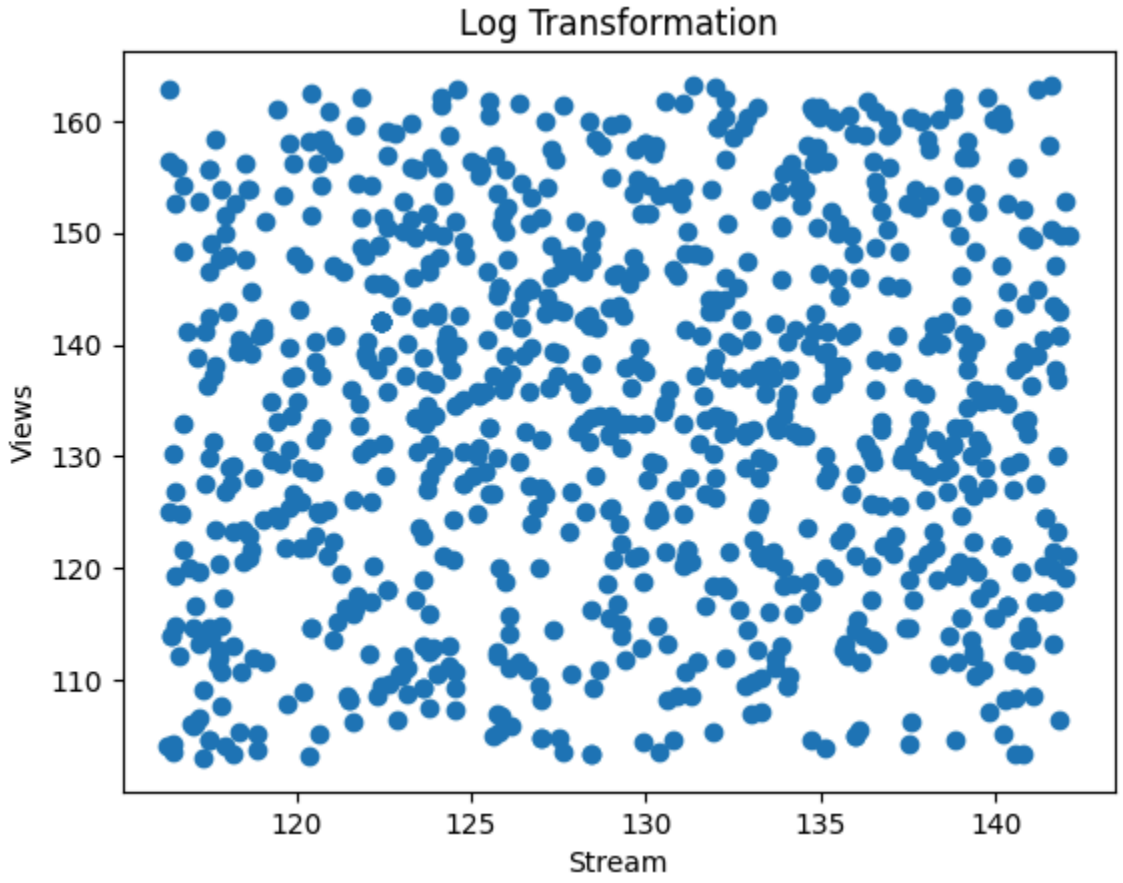
19.



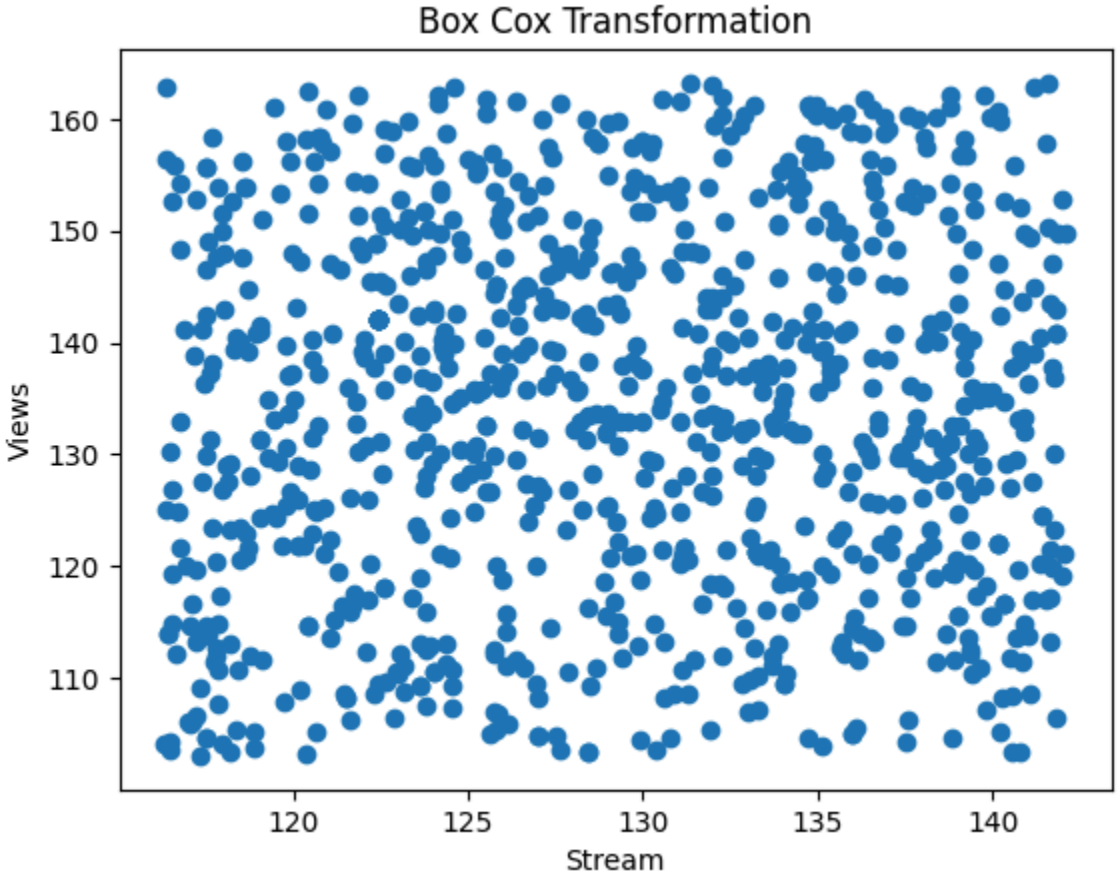
20.


PCA: Views vs Stream
1e6
8
6
4
2
0
−2
−4
Principal Component 2
−1.5
−1.0
−0.5
0.0
0.5
1.0
1.5
1e7
Principal Component 1

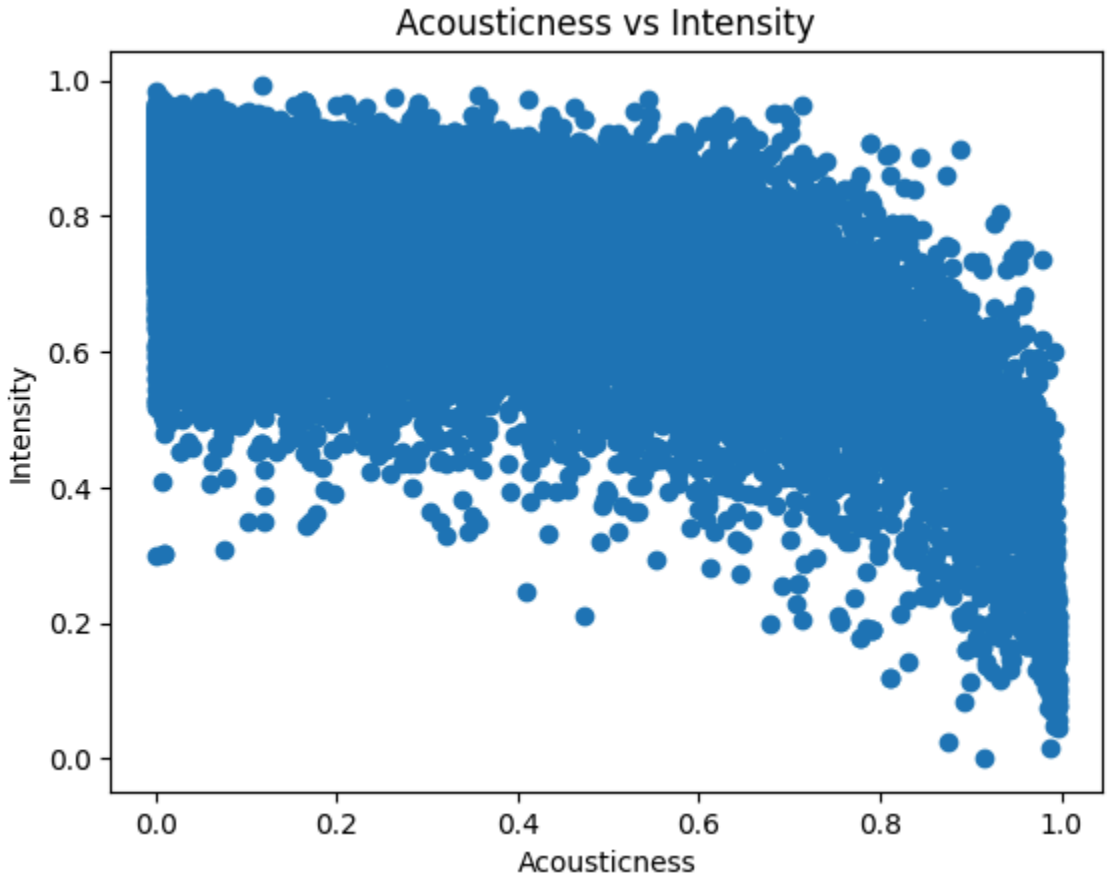
21.


Log Transformation
160
150
140
130
120
110
Views
120
125
130
135
140
Stream

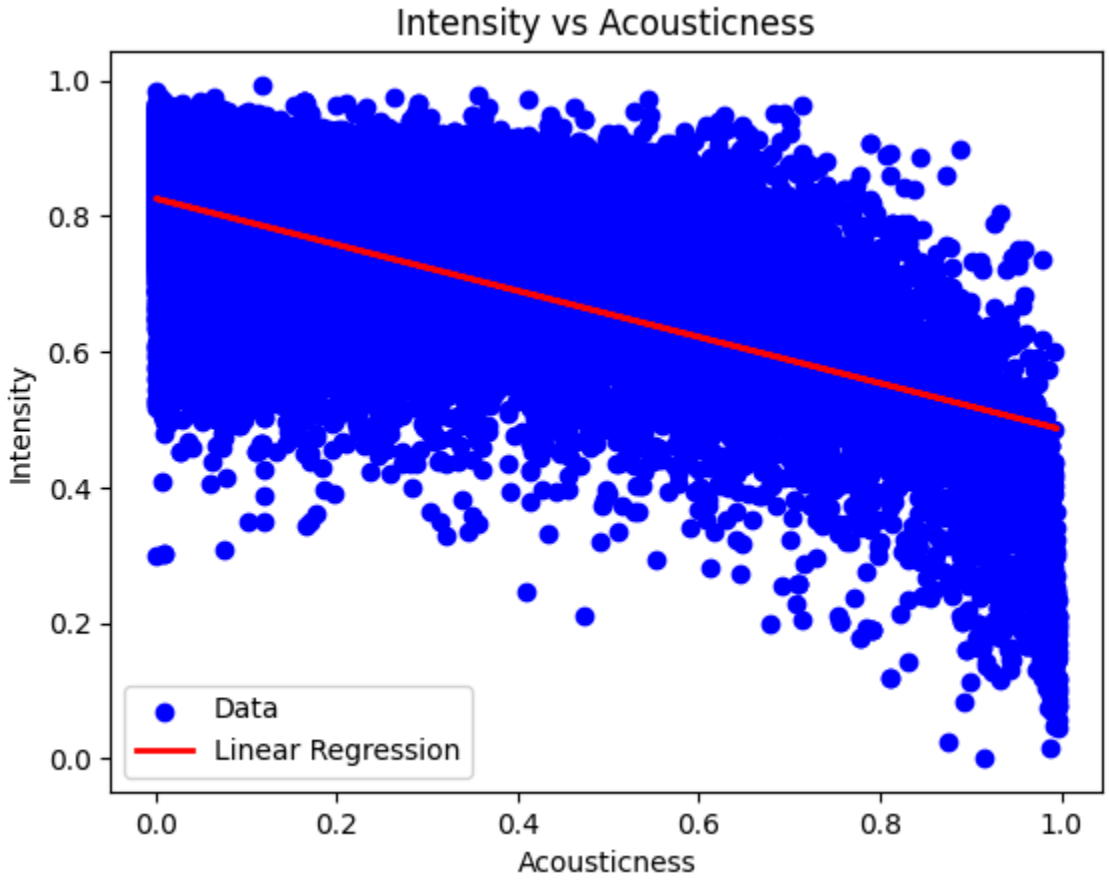
22.


Box Cox Transformation
160
150
140
130
120
110
Views
120
125
130
135
140
Stream

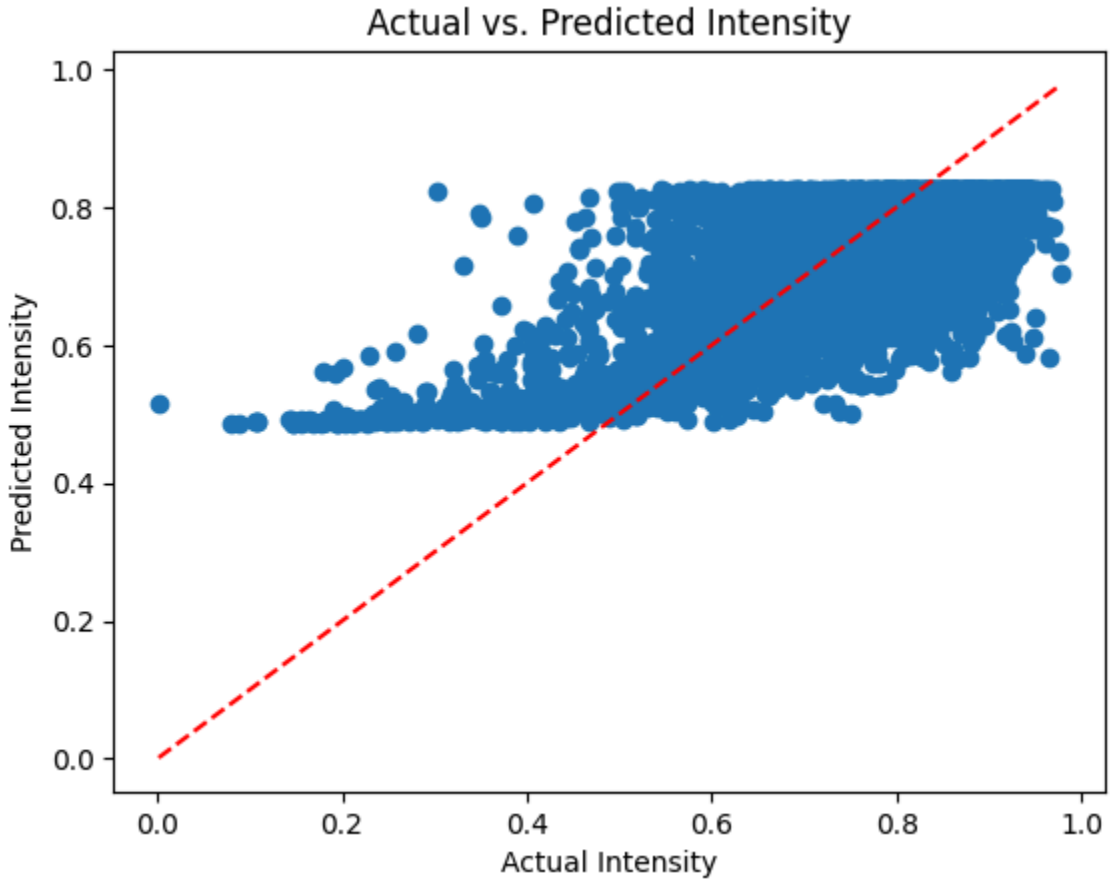
23.


Acousticness vs Intensity
Intensity
Acousticness
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0

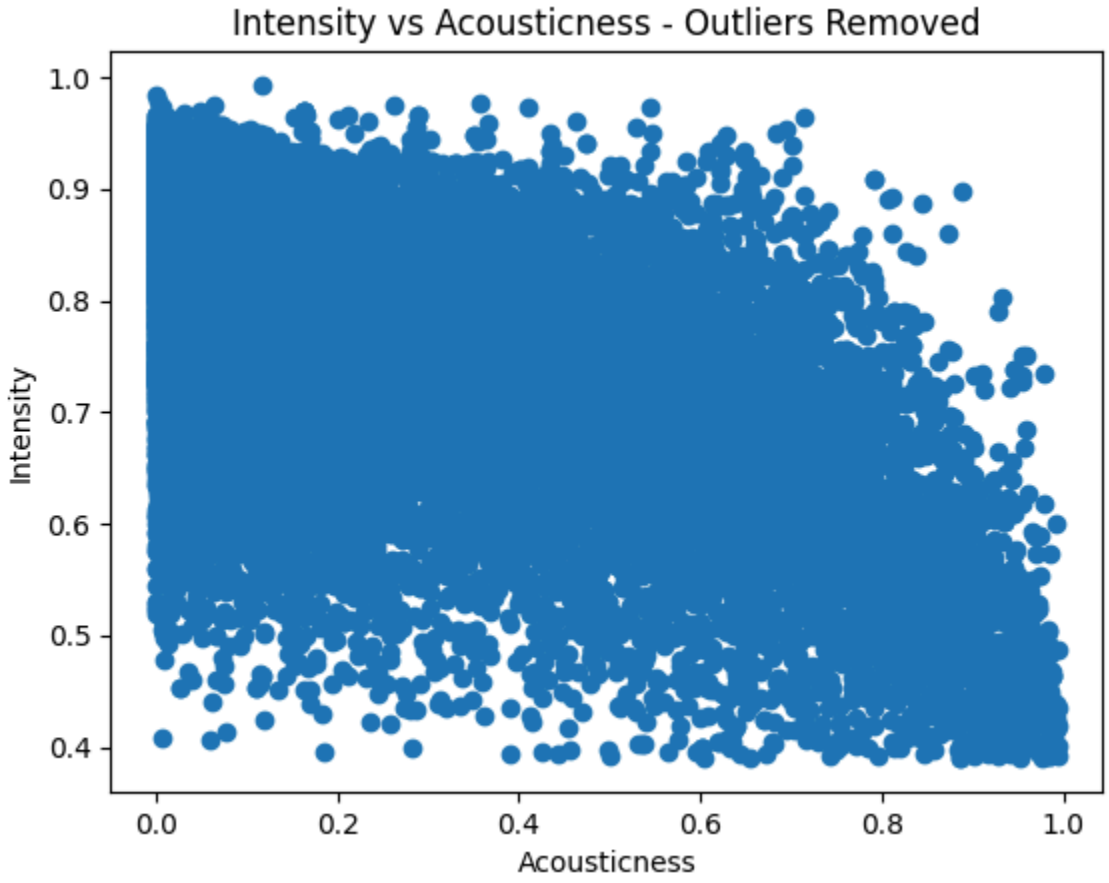
24.


Intensity vs Acousticness
Intensity
Acousticness
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0
Data
Linear Regression

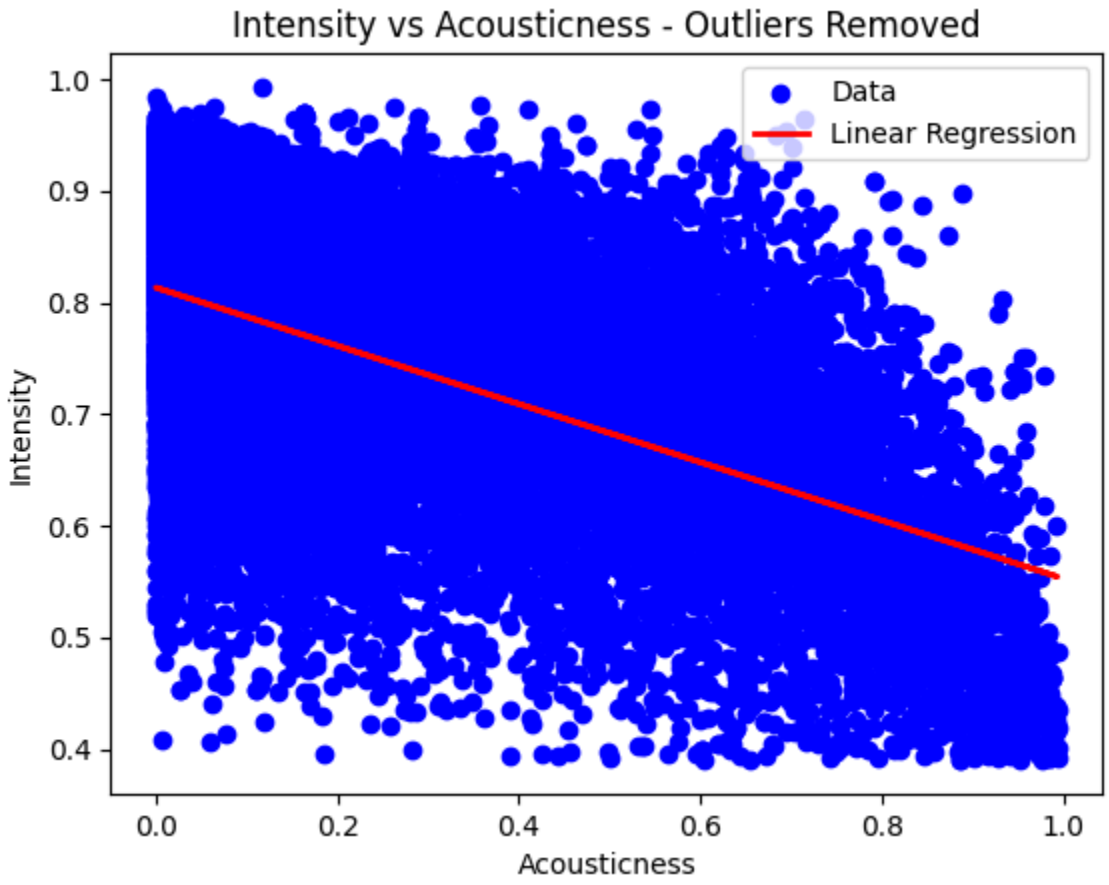
25.


Actual vs. Predicted Intensity
Predicted Intensity
Actual Intensity
0.0
0.2
0.4
0.6
0.8
1.0

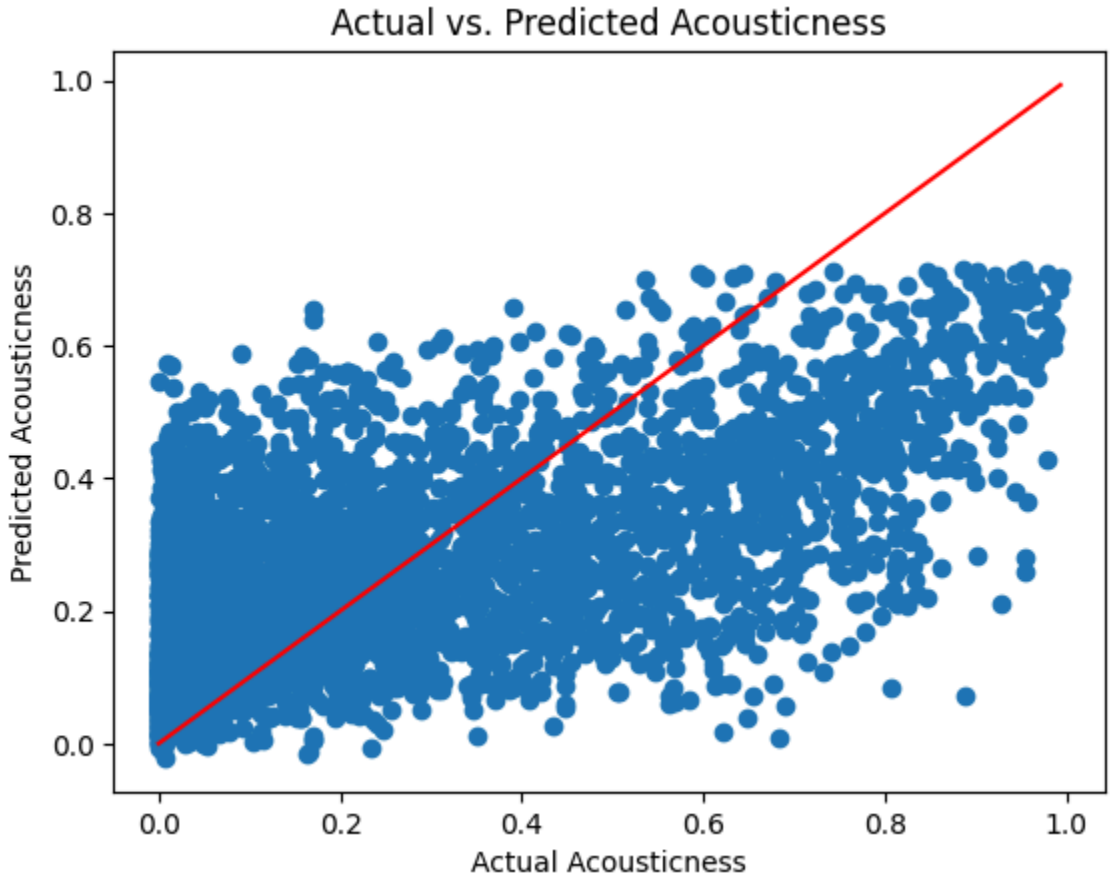
26.


Intensity vs Acousticness - Outliers Removed
Intensity
Acousticness
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.0
0.2
0.4
0.6
0.8
1.0

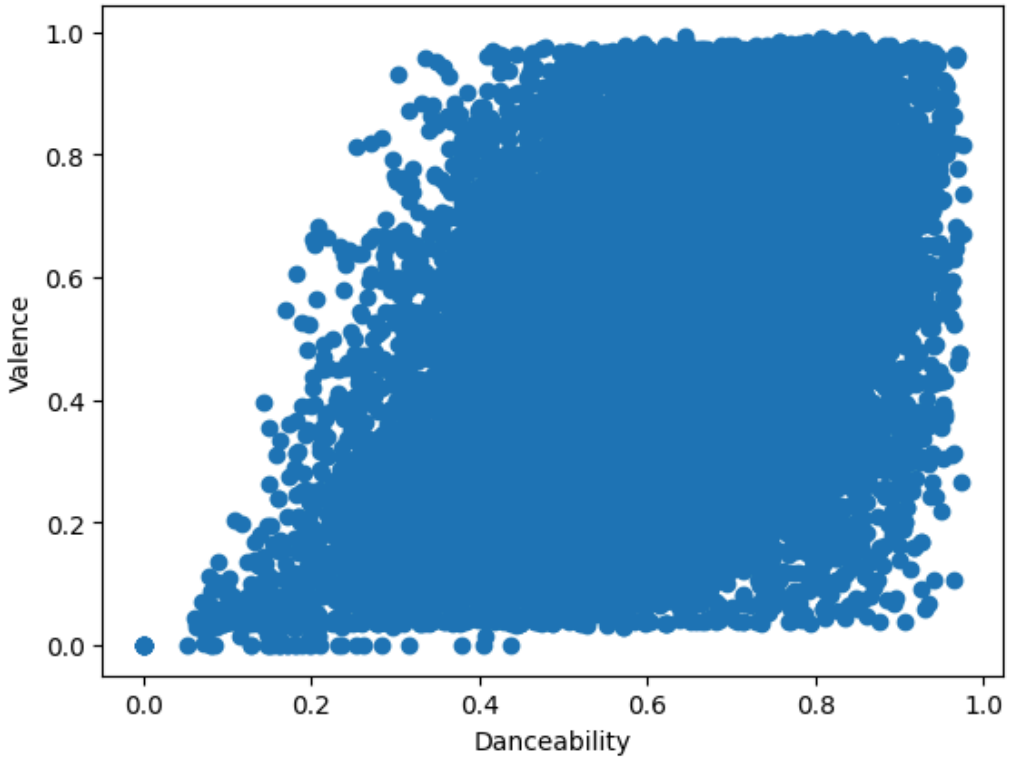
27.


Intensity vs Acousticness - Outliers Removed
Data
Linear Regression
Intensity
Acousticness
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.0
0.2
0.4
0.6
0.8
1.0

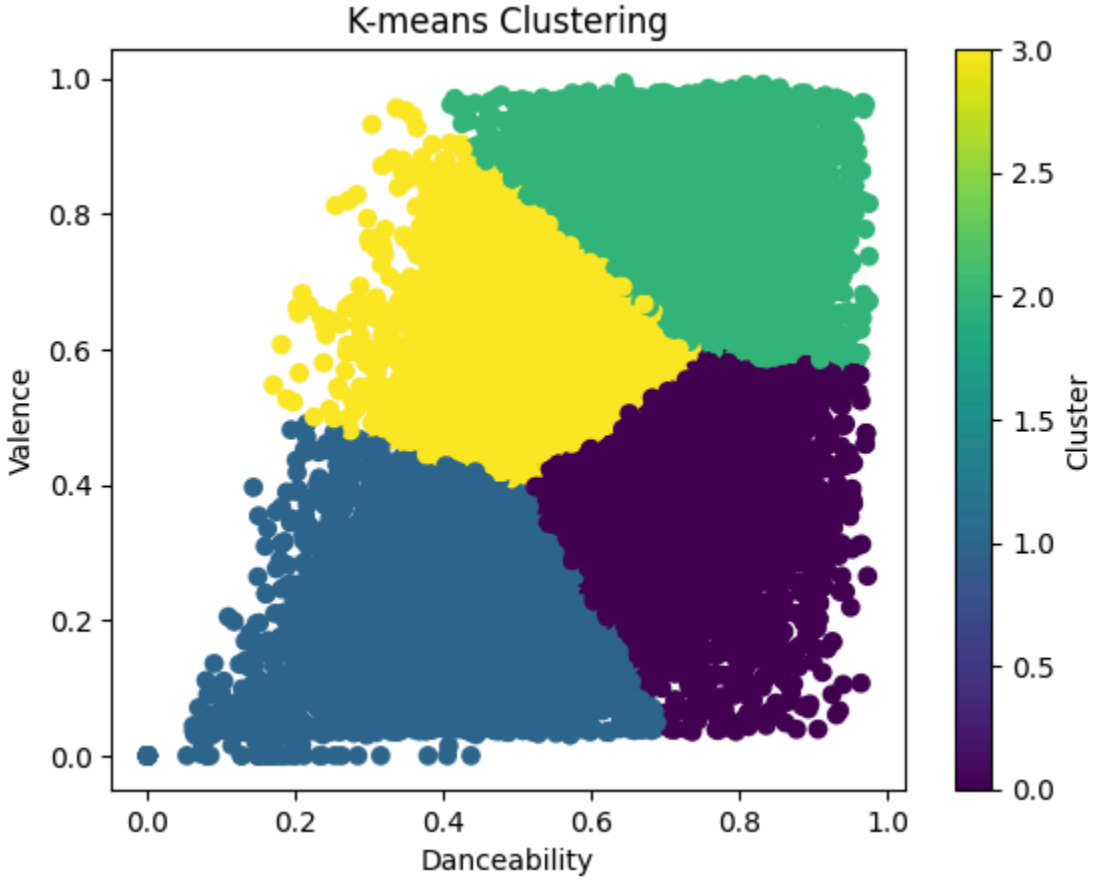
28.


Actual vs. Predicted Acousticness
Predicted Acousticness
Actual Acousticness
0.0
0.2
0.4
0.6
0.8
1.0

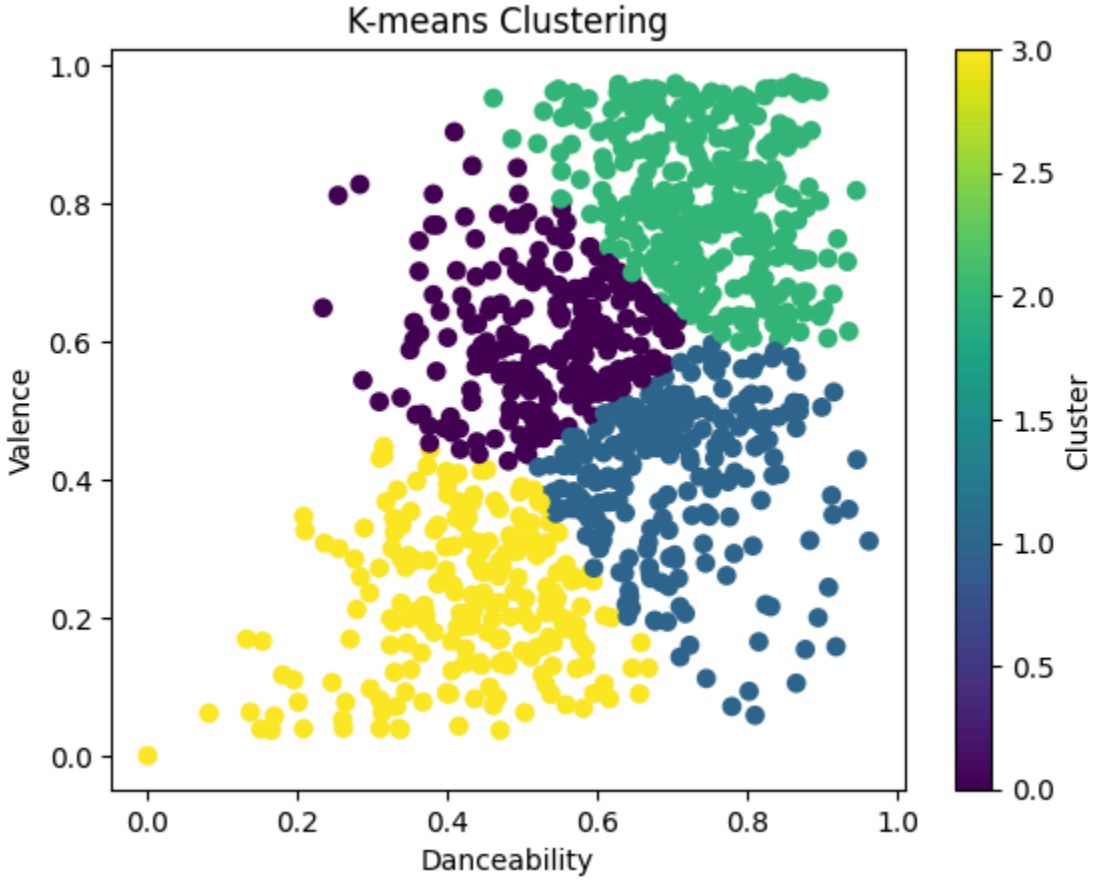
29.



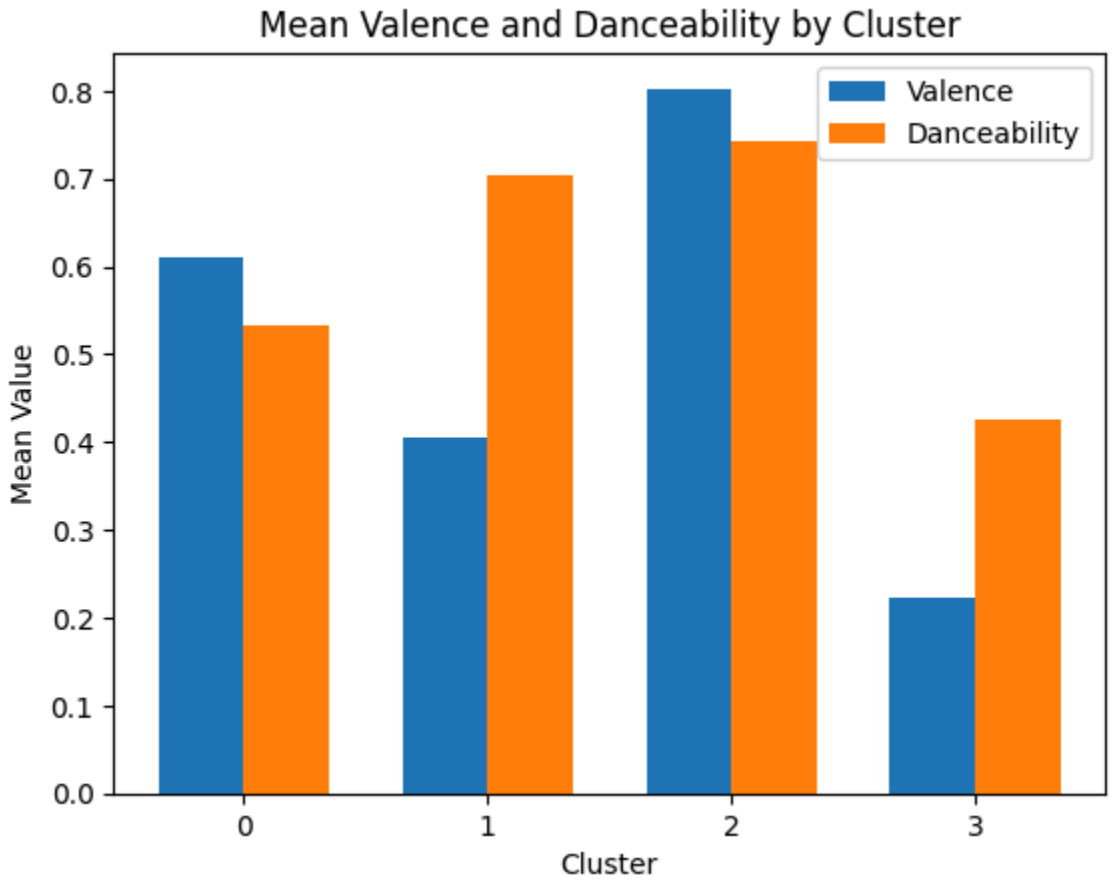
30.


K-means Clustering
Valence
Danceability
Cluster
0.0
0.2
0.4
0.6
0.8
1.0
0.5
1.5
2.0
2.5
3.0

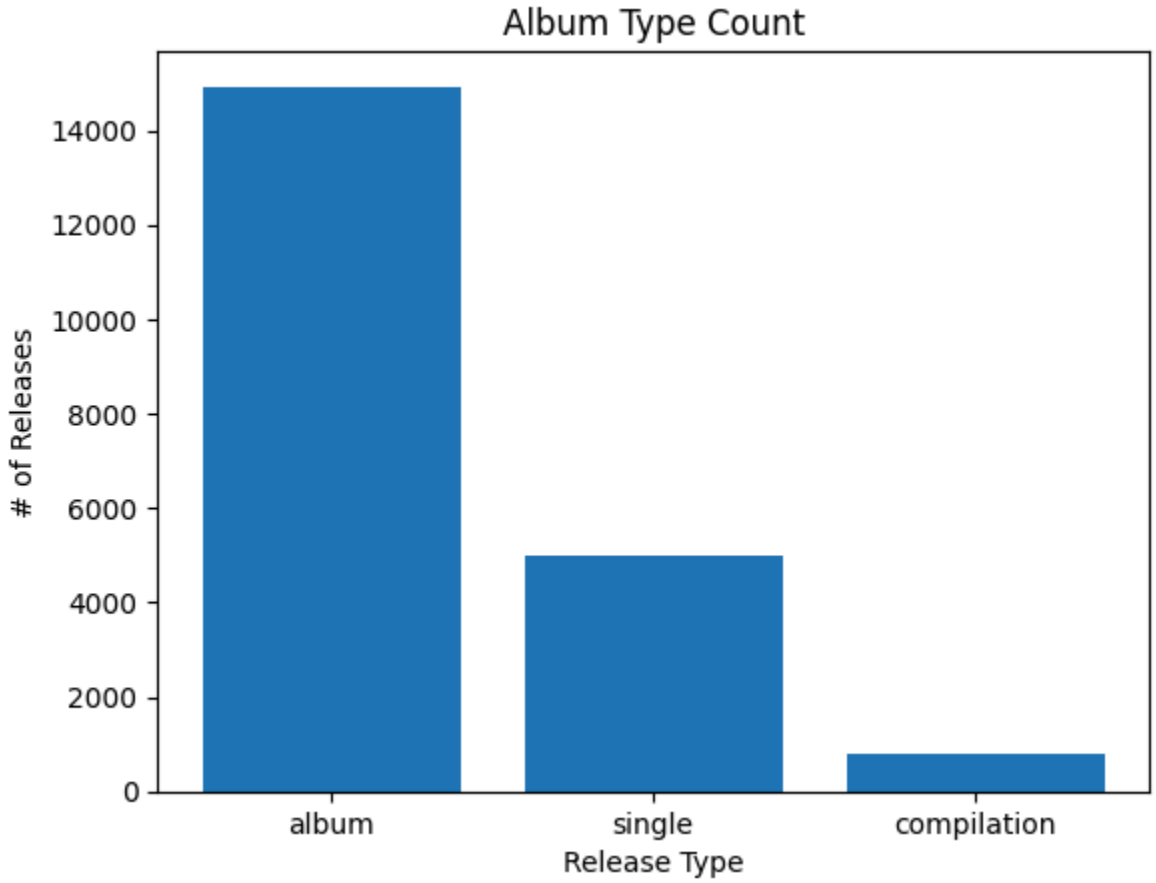
31.


K-means Clustering
Valence
Danceability
Cluster
0.0
0.2
0.4
0.6
0.8
1.0
0.5
1.5
2.0
2.5
3.0

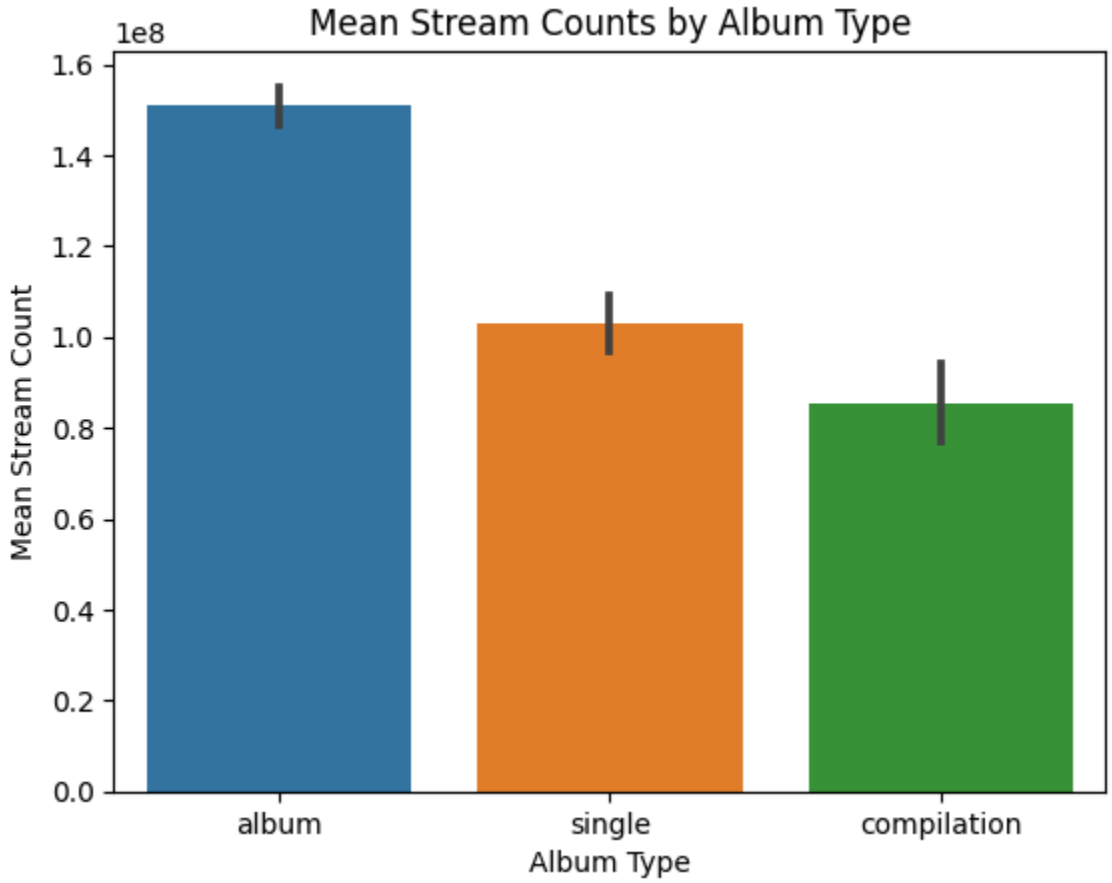
32.


Mean Valence and Danceability by Cluster
Valence
Danceability
Mean Value
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0
1
2
3
Cluster

33.


Album Type Count
# of Releases
14000
12000
10000
8000
6000
4000
2000
0
album
single
compilation
Release Type

34.


Mean Stream Counts by Album Type
1e8
1.6
1.4
1.2
1.0
0.8
0.6
0.4
0.2
0.0
Mean Stream Count
album
single
compilation
Album Type