# **Spotify top 10k songs 1960 to 2023**

**Exploratory Analysis**

John Cason, [jcason@bellarmine.edu](mailto:jcason@bellarmine.edu)

Sawyer Mason, [emason2@bellarmine.edu](mailto:emason2@bellarmine.edu)

1. **INTRODUCTION**

The Data set we chose to use for this project was Spotify top 10k songs from 1960-2023. This data set was found on Kaggle and was made to help students increase their familiarity with R, Tableau, and Jupiter Notebook while investigating some interesting data on popular songs within Spotify. This list of 10k songs is from Spotify so it has a limited catalog to songs provided by Spotify and is missing some major artists and songs. We chose to use this because of its large variety of categories, data types, and variables. This data set also seemed very intriguing. Spotify gears all its listeners to kid friendly music so there is not fully accurate explicit data. These things are fine, but they're worth keeping in mind while we look at relationships. Also, remember that this data set has a limited database to what's provided on Spotify, so these relationships are not a full representation of music.

1. **DATA SET DESCRIPTION**

This data set contains 10,000 samples with 29 columns with 1 bool, 1 datetime, 12 float64, 4 int64, and 11 object data types. Most of the object variables are entirely unique per entry A complete listing is shown in **Table 1**.

**Table 1: Data Types and Missing Data :**

**Data in place after dropping certain columns and the “Year” column**

|  |  |  |
| --- | --- | --- |
| *Variable Name* | *Data Type* | *Missing Data (%)* |
| Track Name | Object | 0.01% |
| Artist Name(s) | Object | 0.01% |
| Album Name | Object | 0.01% |
| Album Artist Name(s) | Object | 0.02% |
| Album Release Date | Object | 0.02% |
| Disc Number | Nominal-Int64 | 0% |
| Track Number | Nominal-Int64 | 0% |
| Track Duration (ms) | Nominal-Int64 | 0% |
| Explicit | Boolean | 0% |
| Popularity | Interval-Int64 | 0% |
| ISRC | Object | 0.03% |
| Added By | Object | 0% |
| Added At | Object | 0% |
| Artist Genres | Object | 5.5% |
| Danceability | Nominal-Float64 | 0.02% |
| Energy | Nominal-Float64 | 0.02% |
| Key | Nominal-Float64 | 0.02% |
| Loudness | Nominal-Float64 | 0.02% |
| Mode | Nominal-Float64 | 0.02% |
| Speechiness | Nominal-Float64 | 0.02% |
| Acousticness | Nominal-Float64 | 0.02% |
| Instrumentalness | Nominal-Float64 | 0.02% |
| Liveness | Nominal-Float64 | 0.02% |
| Valence | Nominal-Float64 | 0.02% |
| Tempo | Nominal-Float64 | 0.02% |
| Time Signature | Nominal-Float64 | 0.02% |
| Label | Object | 0.06% |
| Copyrights | Object | 0.24% |
| Year | Interval-Datetime | 0.03% |

1. **Data Set Summary Statistics**

In total, there are 16 variables that can be summarized using standard statistics. There are 4 Int64 variables and 12 Float64 Variables. They describe a range of criteria for each entry such as the duration of the song and its unique identifiers. In addition, it also describes many of the technical aspects of the song such as loudness and tempo.

**Table 2: Summary Statistics for Spotify top 10k songs 1960 to 2023**

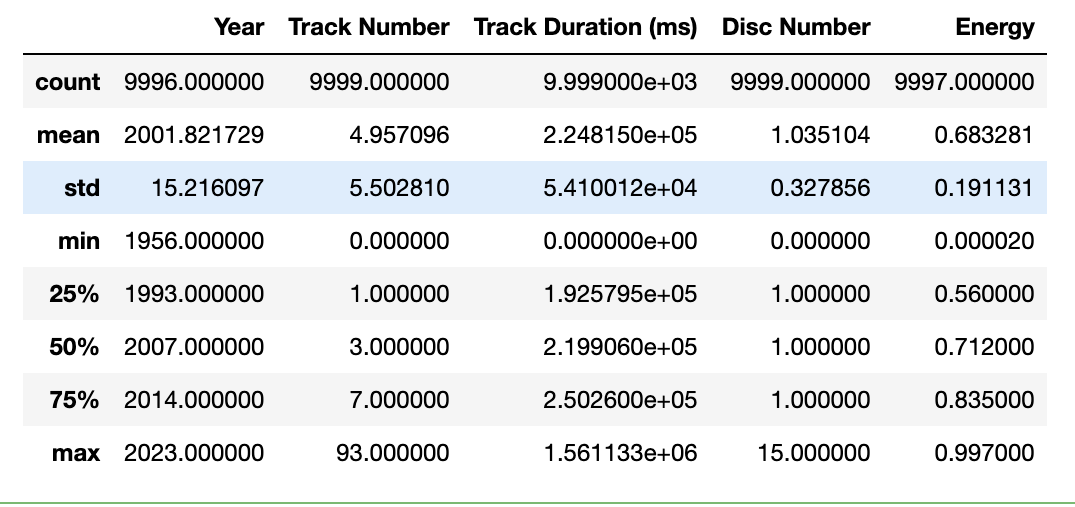
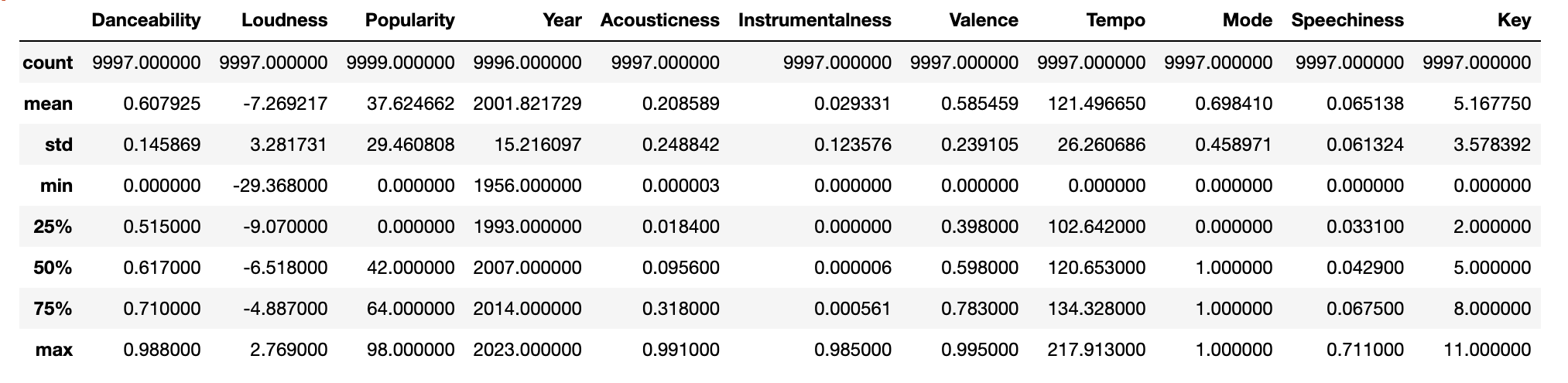


Table 3: Proportions for Spotify top 10k songs 1960 to 2023 (n=9,999)

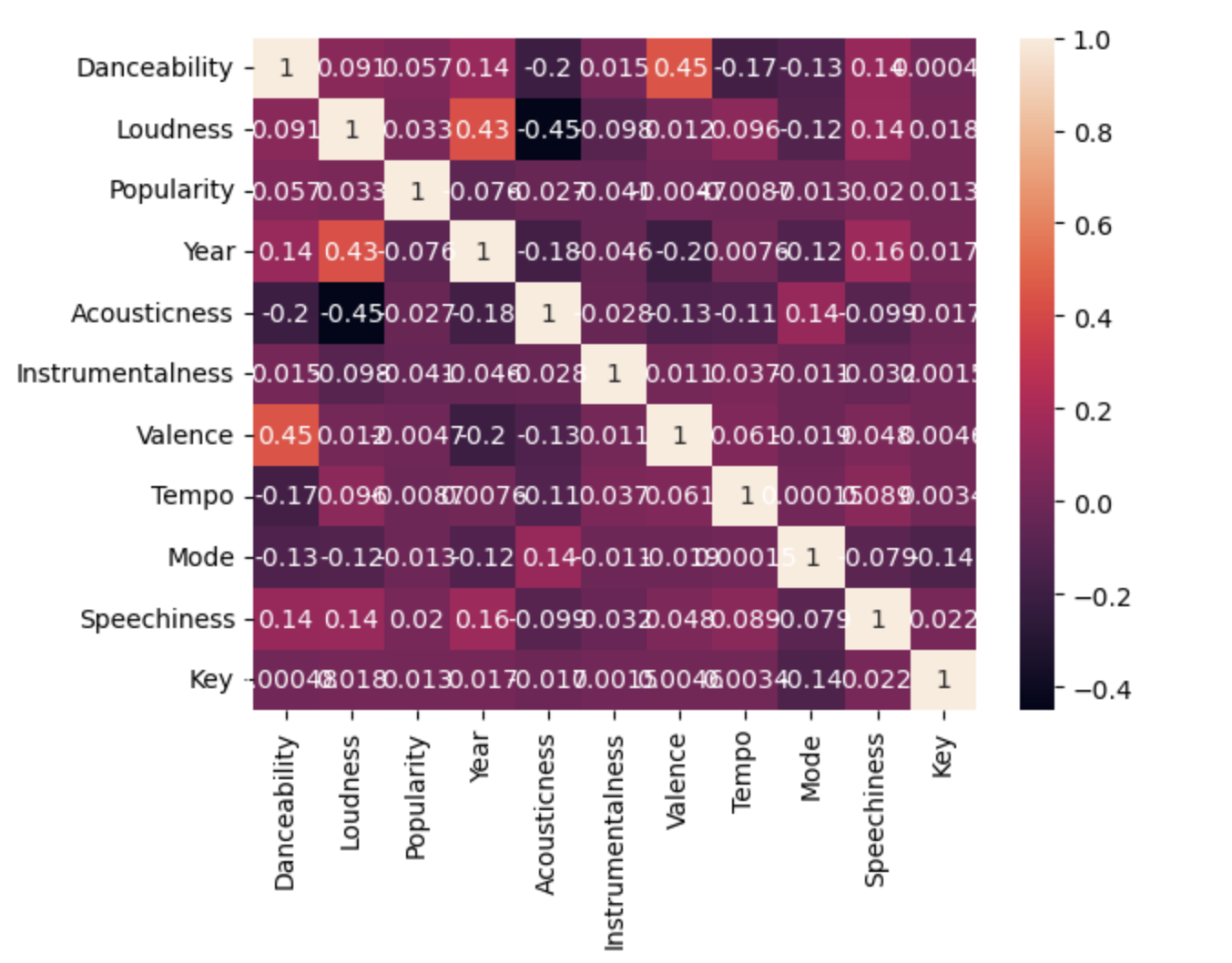
Due to not having any “traditional” categorical variables makes this area of analysis very difficult. Most of the categorical variables are almost entirely unique and thus unrealistic and not helpful to individually list in a data set of this size. As a result, we have chosen to omit them.

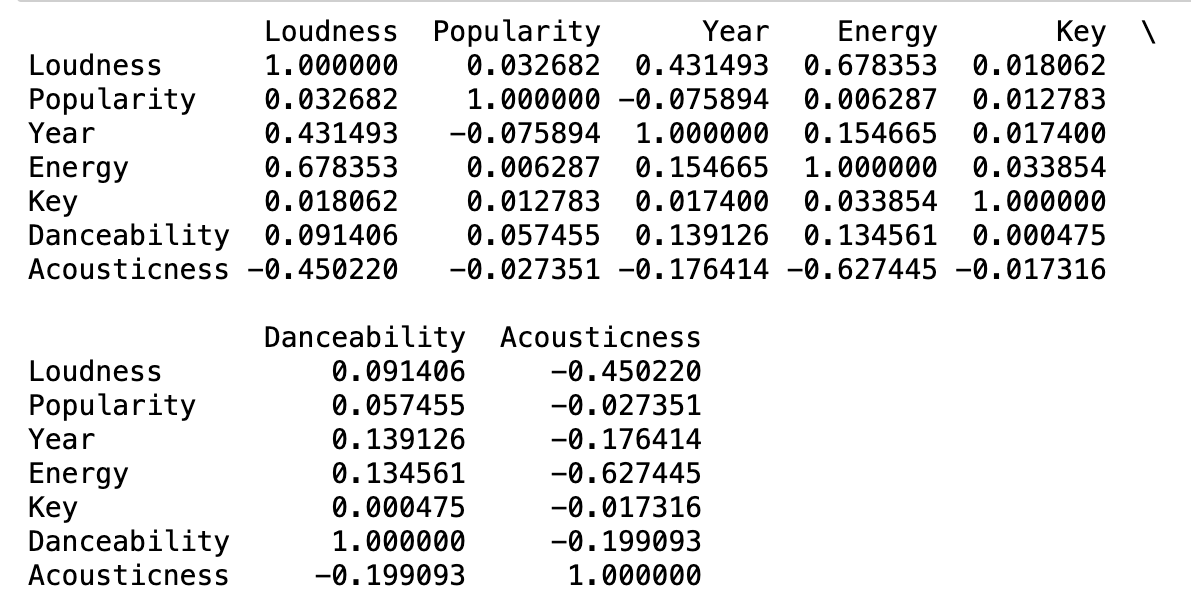
|  |  |  |
| --- | --- | --- |
| *Category* | *Frequency* | *Proportion (%)* |
| *Explicit* | *9,999* | *True 509/5.09%, False: 9,490/94.91%* |

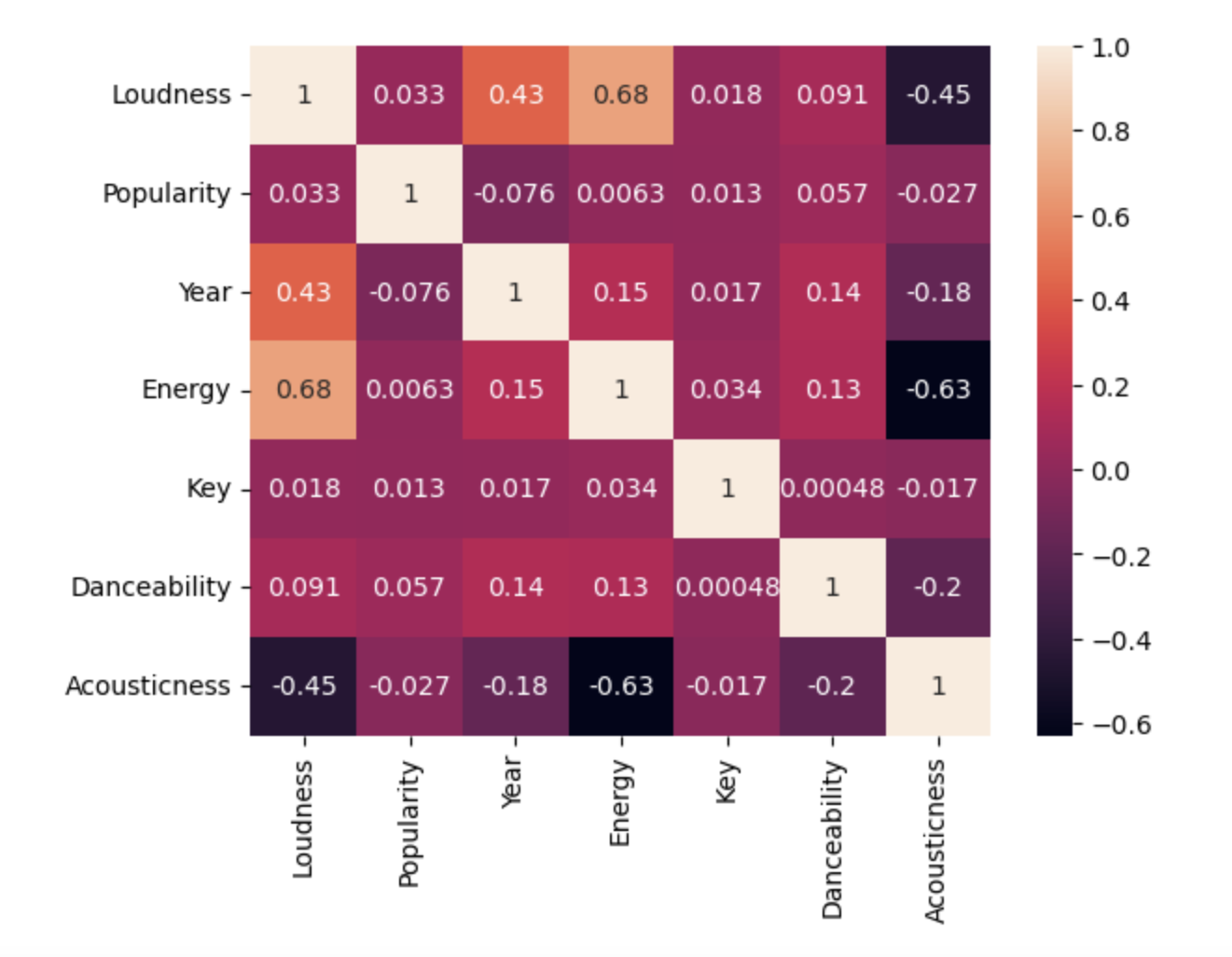
Table 4: Correlation Table/Tables

A screenshot of a computer

Description automatically generated







1. **DATA SET GRAPHICAL EXPLORATION**

Initially, we graphically focused on the popularity metric to decern what the most popular songs/song types that were present in the data set. This incudes a bar chart of the top 10 most popular genres throughout the years (Figure 1), the top 10 most popular artists in the data set by year they were popular (Figure 2), and the most numerous artist genres in the data set (Figure 3). The main categorical variable that was analyzed was whether a song was considered “Explicit” or not. The count of explicit songs per year can be seen in the through the line chart in Figure 4. Finally, by using the heat map above as a reference, we utilized scatter plots to display the correlations between Loudness vs Acousticness(Figure 5) and Loudness vs Energy (Figure 6).

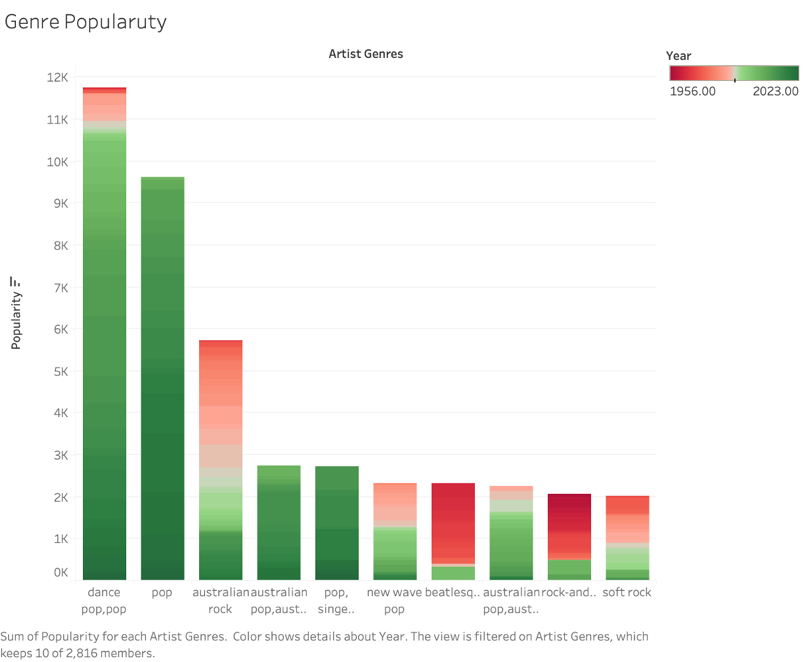


Figure 1: Bar Chart of top 10 Genre Popularity colored by Year

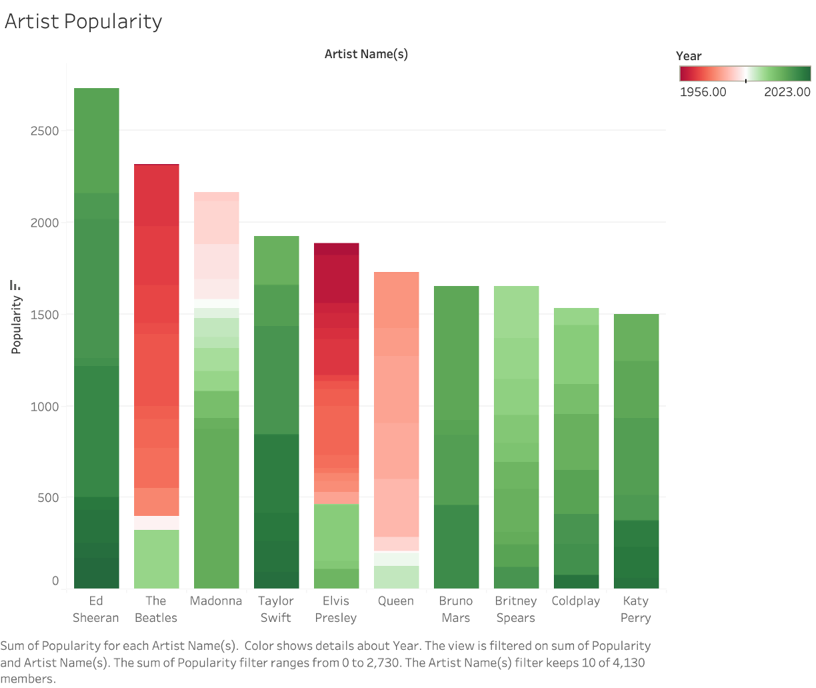


Figure 2: Bar Chart of top 10 Artist Popularity Colored by Year

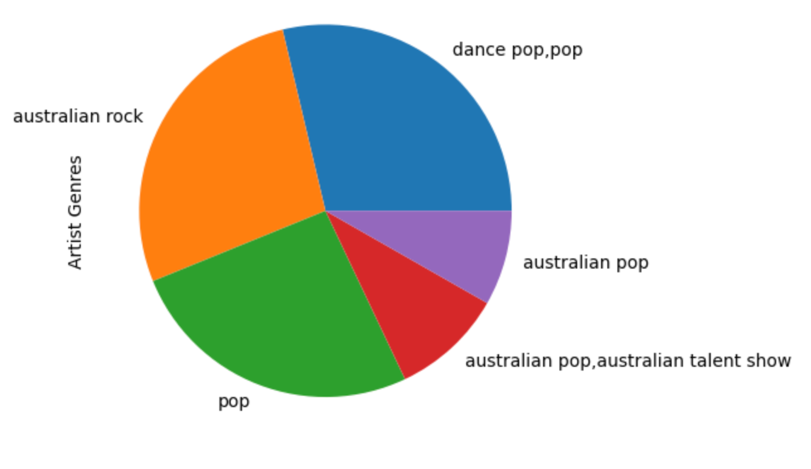


Figure 3: Pie Chart of Top 5 all time most frequent Artist Genres

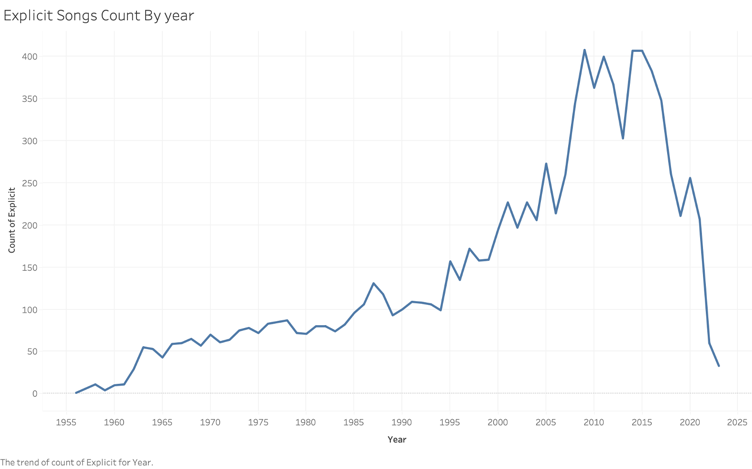


Figure 4: Line plot of Count of explicit songs per year

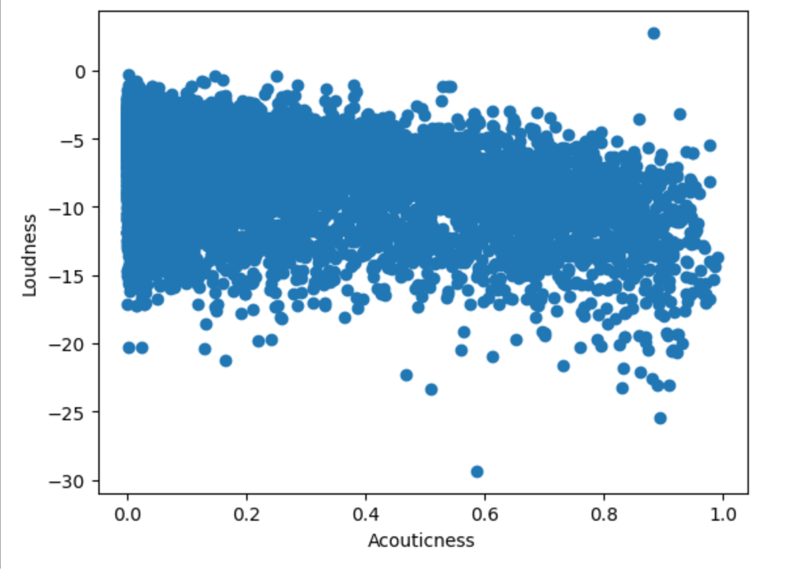


Figure 5: loudness vs Acousticness Scatter Plot

Shows a weak negative correlation between Loudness and Acousticness which is further proven by the Heat map

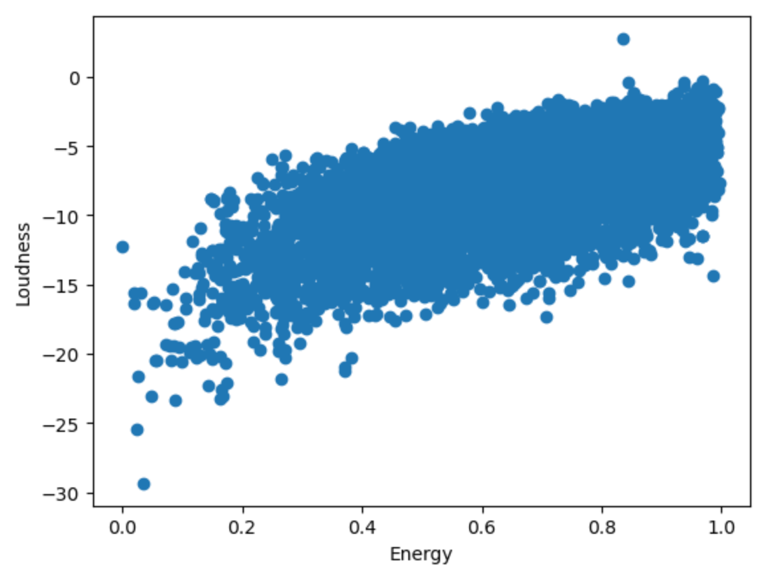


Figure 6: Loudness vs Energy Scatter Plot

Shows a strong positive correlation between Loudness and Energy which is further proven by the Heat map

1. **SUMMARY OF FINDINGS**

Through the use of the heat mat we found that no numerical data present had any statistically significant correlation to the popularity of a given song. Thus, we concluded that artist, and artist genre, unique variables, had the greatest overall effect on popularity. The two most popular being rock and pop, which the most popular artist, Ed Sheeran, falls into. In addition, we found that overtime, loudness and energy had a positive correlation while loudness and acousticness had a negative correlation, thus reinforcing idea that rock and pop have become more and more popular. The rise of “explicit” songs also provides interesting data as it is trending upwards. Thus, taken with the previous findings, one can conclude that louder, more energetic, and typically explicit songs tend to propel an artist and song into to top popular rankings if they are in certain genres of music.