Data Science 1st Assignment

Dataset: Top Songs on Spotify

Mahan Madani – 99222092

1. Overview

This dataset is comprised of the top 10000 songs on Spotify, based on the data from ARIA and Billboard Charts. It includes over 60 years of tracking; from 1960 up until July 2023.

The dataset spans multiple music genres and depicts the growth of musical trends throughout time, providing insights into the ever-changing environment of popular music.

2. Dataset Exploration & Analysis

The dataset stores 35 different attributes relating to each song. Some of the columns are irrelevant and therefore will be omitted from this report.

Below you can find a list of some of the key attributes:

- **"Track Name":** The Title of the song. (string)
- **"Artist Name(s)":** The names of all the artists featured in a song. Each name is separated by a comma. (string)
- **"Album Name":** The title of the album the song belongs to. (string)
- **"Album Artist Name(s)":** The names of all the artists who own the album. Each name is separated by a comma. (string)
- **"Album Release Date":** The official date of the album's release. (date)
- **"Disc Number":** A number indicating which disc the song can be found on. (integer)
- **"Track Number":** A number indicating the song's placement on the album tracklist. (integer)

- **"Track Duration (ms)":** The length of time the song plays from start to finish. (integer)
- **"Explicit":** Indicates whether or not a song contains explicit language. (bool)
- **"Popularity":** A numerical value used to measure the song's popularity. (integer)
- **"ISRC":** The International Standard Recording Code is a 12–14-digit code that identifies a specific recording of a song. (string)
- **"Added At":** The date the song was added to Spotify. (date)
- **"Artist Genres":** A list of genres separated by a comma. (string)
- **"Danceability":** A measure of the song's danceability. (float)
- **"Energy":** A measure of the song's energy level. (float)
- **"Key":** A value indicating what key the song was written in. (float)
- **"Loudness":** A measure of the song's loudness. (float)
- **"Mode":** A binary value indicating the song's mode. (float)
- **"Speechiness":** A measure of the song's Speechiness. (float)
- **"Acousticness":** A measure of the song's Acousticness. (float)
- **"Instrumentalness":** A measure of the song's Instrumentalness. (float)
- **"Liveness":** A measure of the song's Livness. (float)
- **"Valence":** A measure of the song's Valence. (float)
- **"Tempo":** A value indicating the song's tempo. (float)
- **"Time Signature":** A Value indicating the song's Time Signature. (float)
- **"Label":** The title of the record label that published the song. (string)

3. Data Preprocessing

3.1 Drop Unnecessary Features:

The following attributes are of little use in this research, as such they should be dropped:

"Disc Number" – "ISRC" – "Label" – "Album Artist Name(s)" – "Added At" - "Track URI" - "Artist URI(s)" - "Album URI" - "Album Artist URI(s)" - "Album Image URL" - "Track Preview URL" - "Added By" - "Copyrights" - "Album Genres"

*Note: The "Album Genres" column was completely empty in the dataset.

3.2 Handle Null Values:

Out of 21 selected features, several contain null values that must be dealt with. The exact number of null values for each attribute is displayed in this table:

	Null Count
Track Name	1
Artist Name(s)	1
Album Name	1
Album Release Date	2
Track Number	0
Track Duration (ms)	0
Explicit	0
Popularity	0
Artist Genres	550
Danceability	2
Energy	2
Key	2
Loudness	2
Mode	2
Speechiness	2
Acousticness	2
Instrumentalness	2
Liveness	2
Valence	2
Tempo	2
Time Signature	2

Since the majority of the null values appear in low quantities, I decided to delete any record with null values in any column other than “Artist Genres”.

The records that do not have a value assigned to their “Artist Genres” will be excluded from any statistical tests and hypotheses that require that attribute.

The number of records missing this specific attribute is too high to manually handle or replace with randomly generated values. Ultimately, the best course of action is to not utilize these records for samples that aim to analyze the genre of music.

3.3 Convert Data Types:

- “Album Release Date” is stored as a string in the dataset. By converting it to a datetime object, it can be used to generate useful features and also allow for sorting the data based on release year.

All of the other attributes in the dataset are using the correct datatype so no additional modification is required.

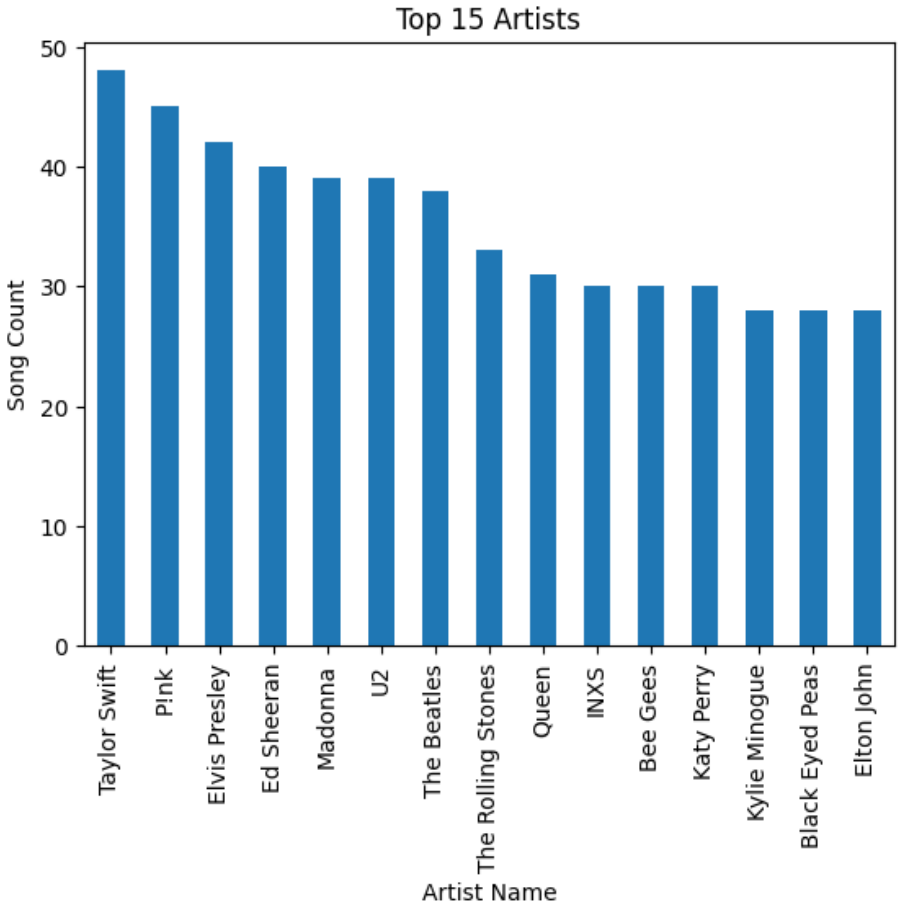
3.4 Feature Generation:

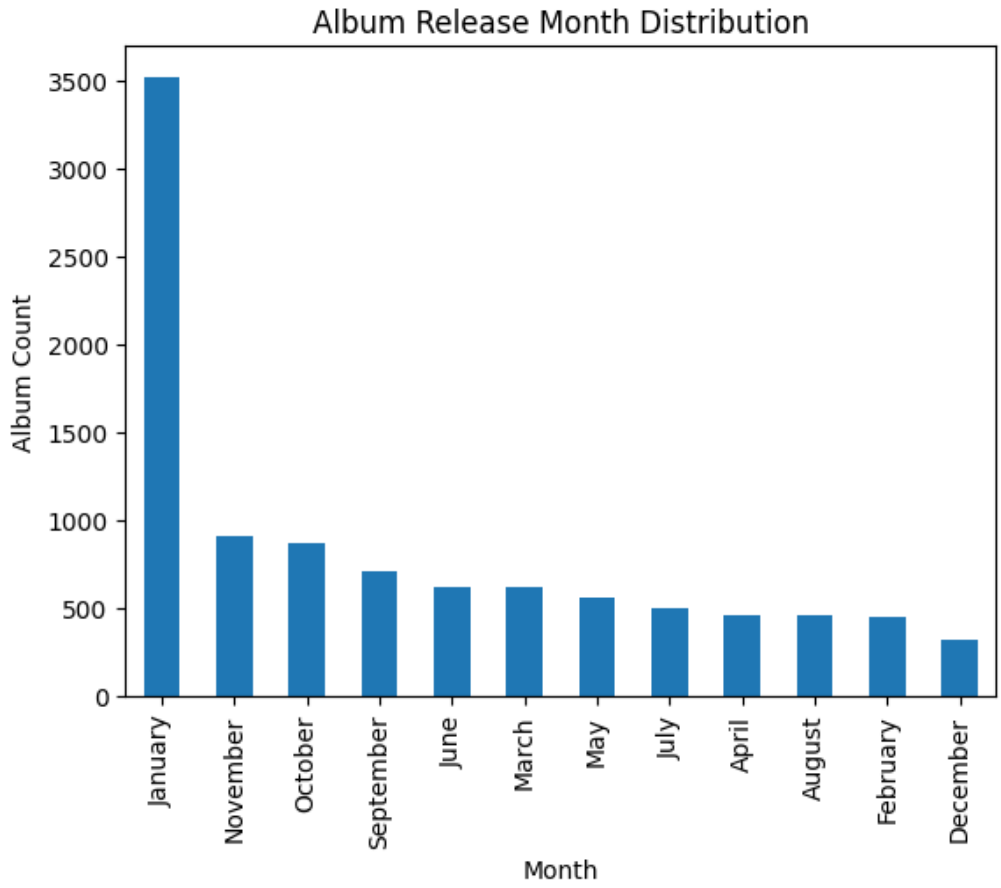
Based on the available data, the following features can be added to further improve the dataset:

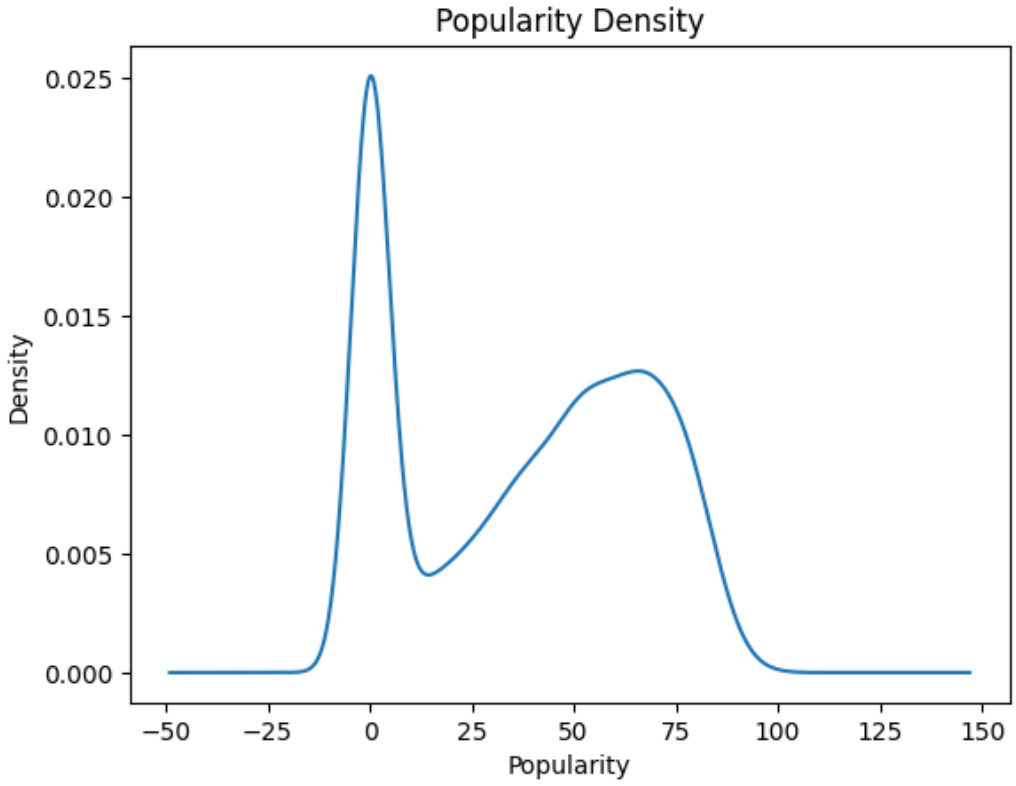
- **Year:** The year the song was released in. Extracted from “Album Release Date”. (integer)
- **Month:** The name of the month the song was released in. Extracted from “Album Release Date”. (string)
- **Track Duration (minutes):** A conversion of the “Track Duration (ms)” attribute to a minutes:seconds format. (timedelta)

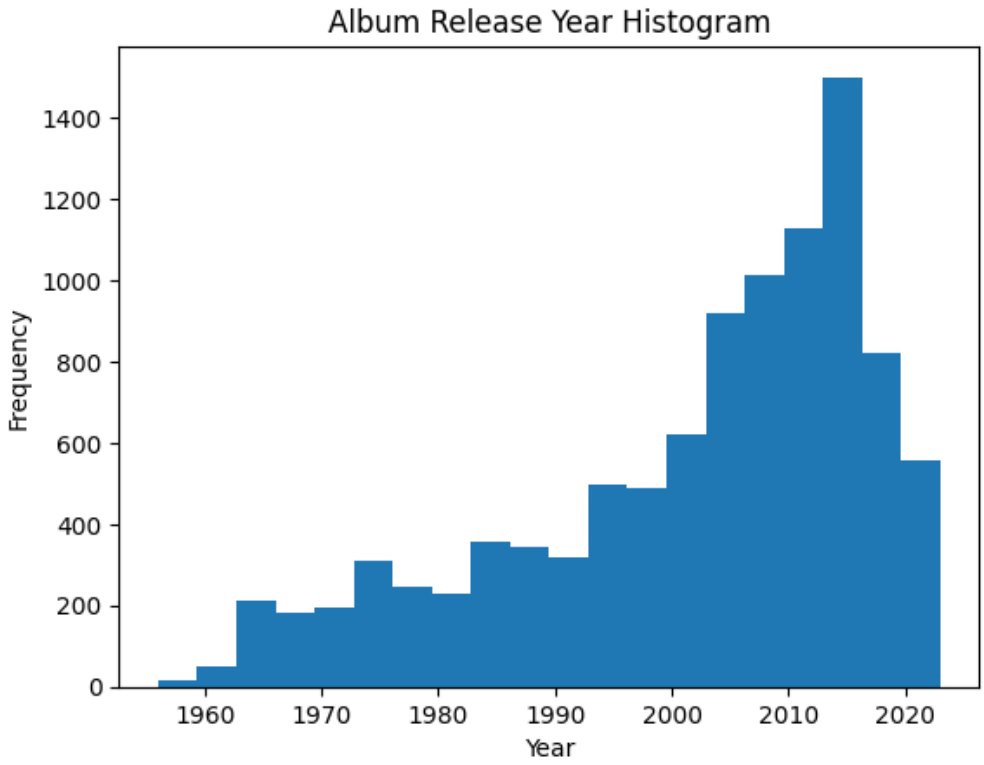
4. Data Visualization:

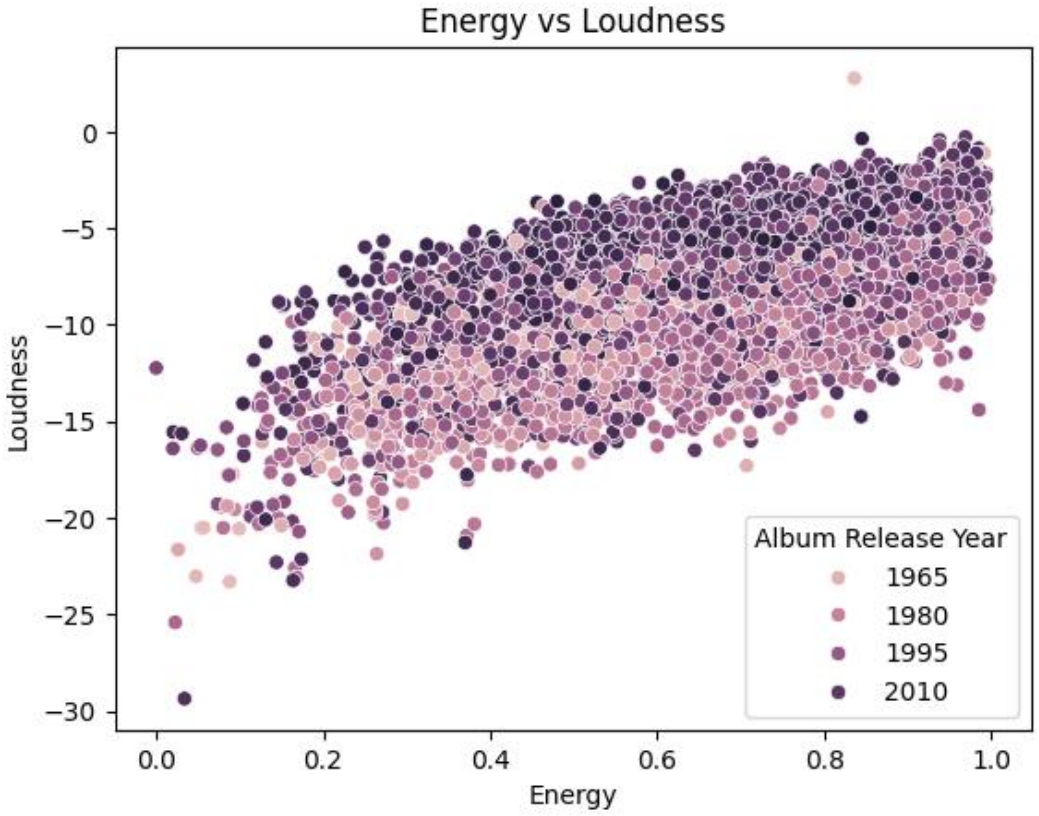
Below you can find a number of plots visualizing the data so that it can be more easily understood. These plots will be used in the next step to conduct a set of statistical tests.

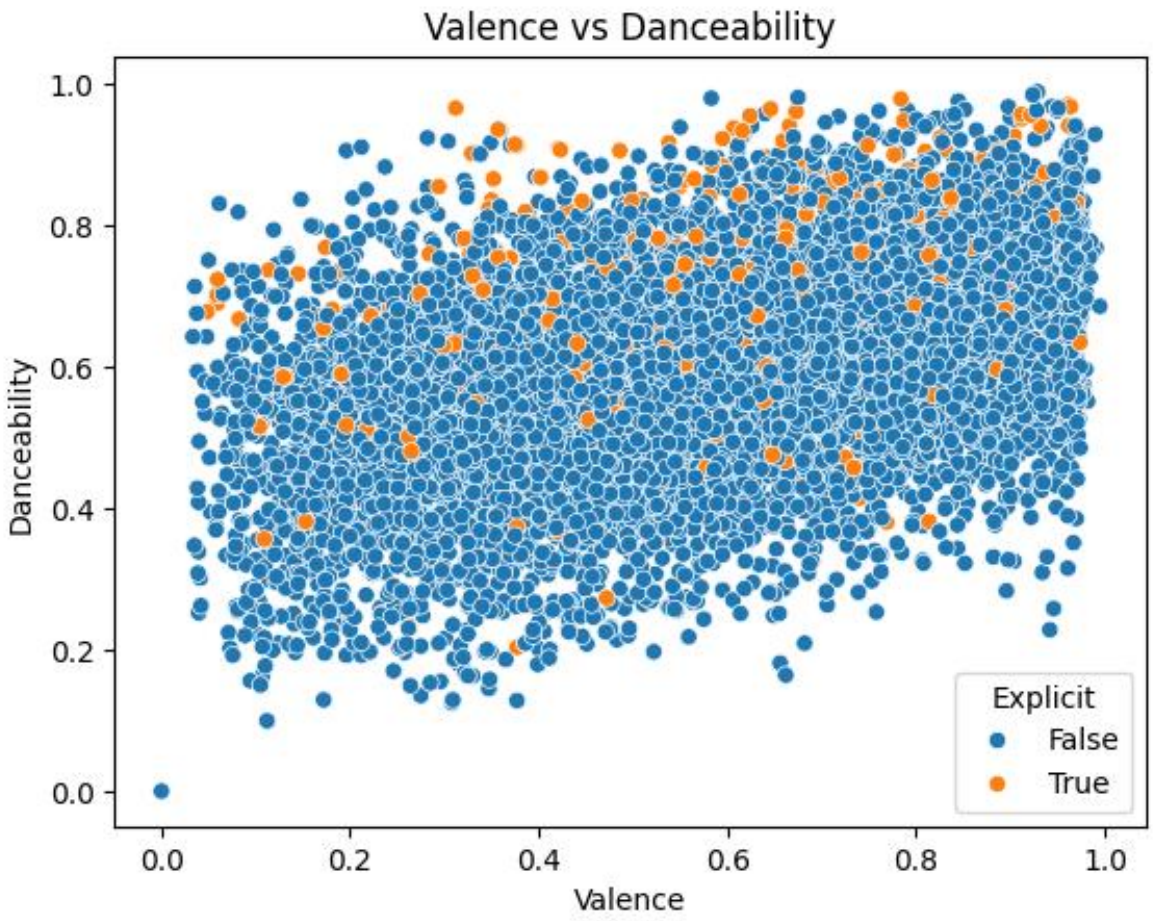


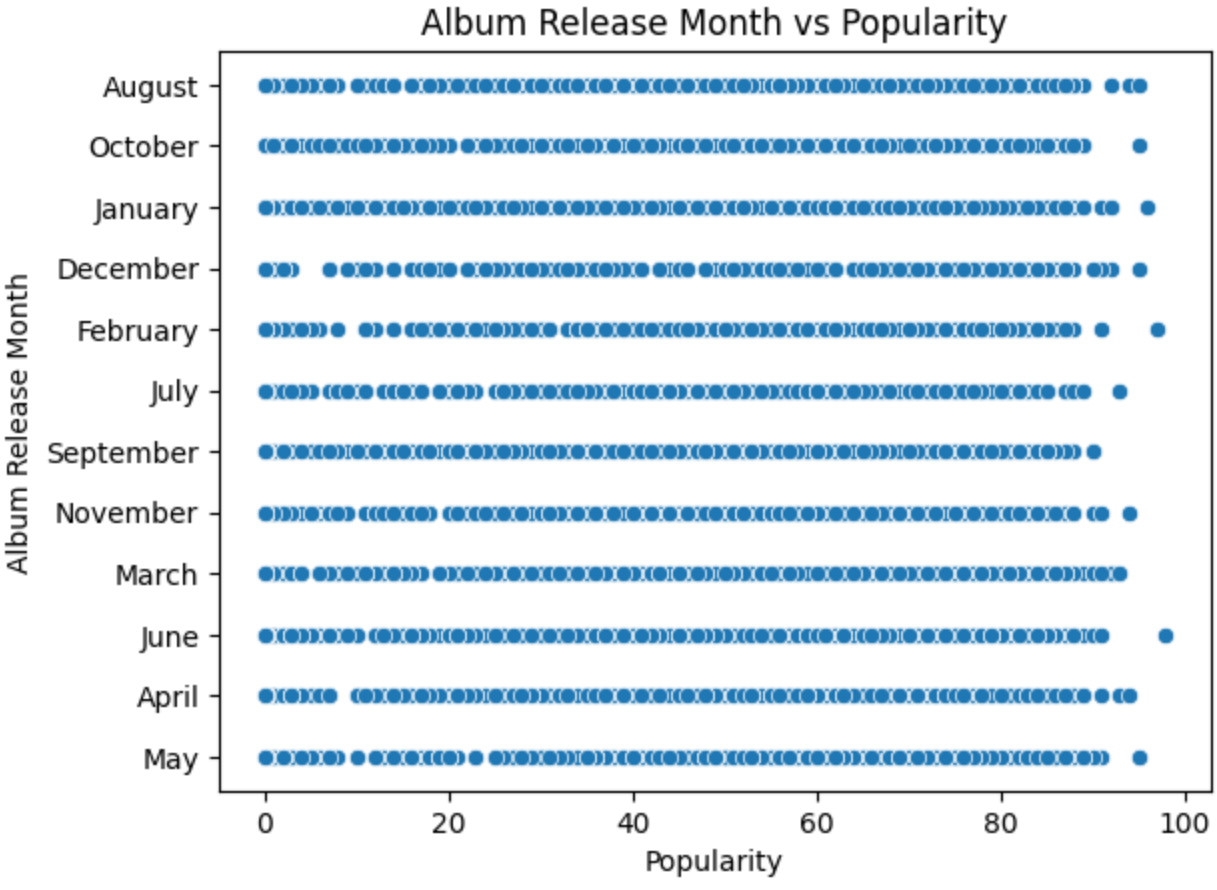

Album Release Month Distribution
Album Count
3500
3000
2500
2000
1500
1000
500
0
January
November
October
September
June
March
May
July
April
August
February
December
Month


Popularity Density
Density
0.025
0.020
0.015
0.010
0.005
0.000
−50
−25
0
25
50
75
100
125
150
Popularity

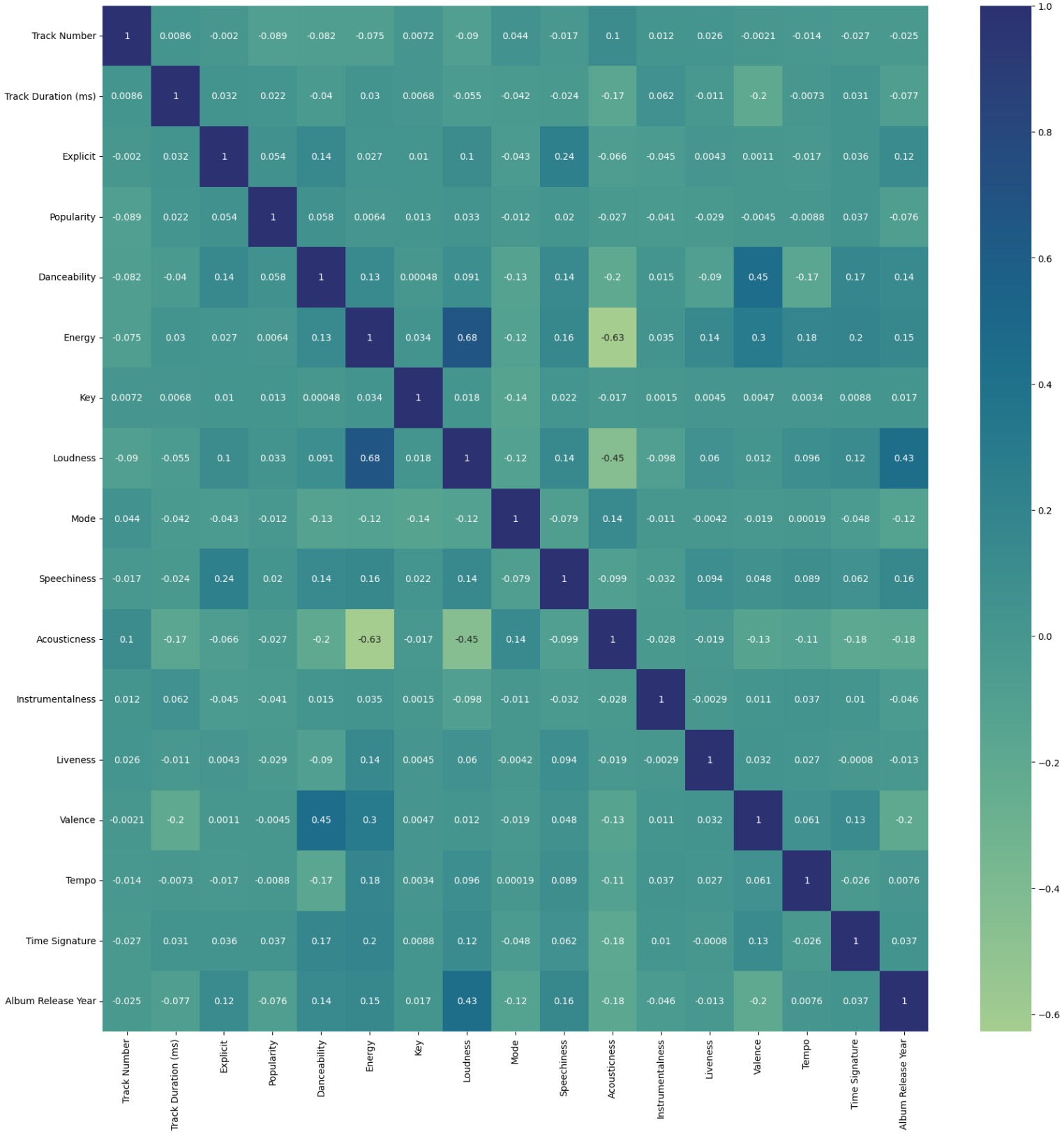

Album Release Year Histogram
Frequency
Year
0
200
400
600
800
1000
1200
1400
1960
1970
1980
1990
2000
2010
2020


Energy vs Loudness
Loudness
Energy
0
−5
−10
−15
−20
−25
−30
0.0
0.2
0.4
0.6
0.8
1.0
Album Release Year
1965
1980
1995
2010

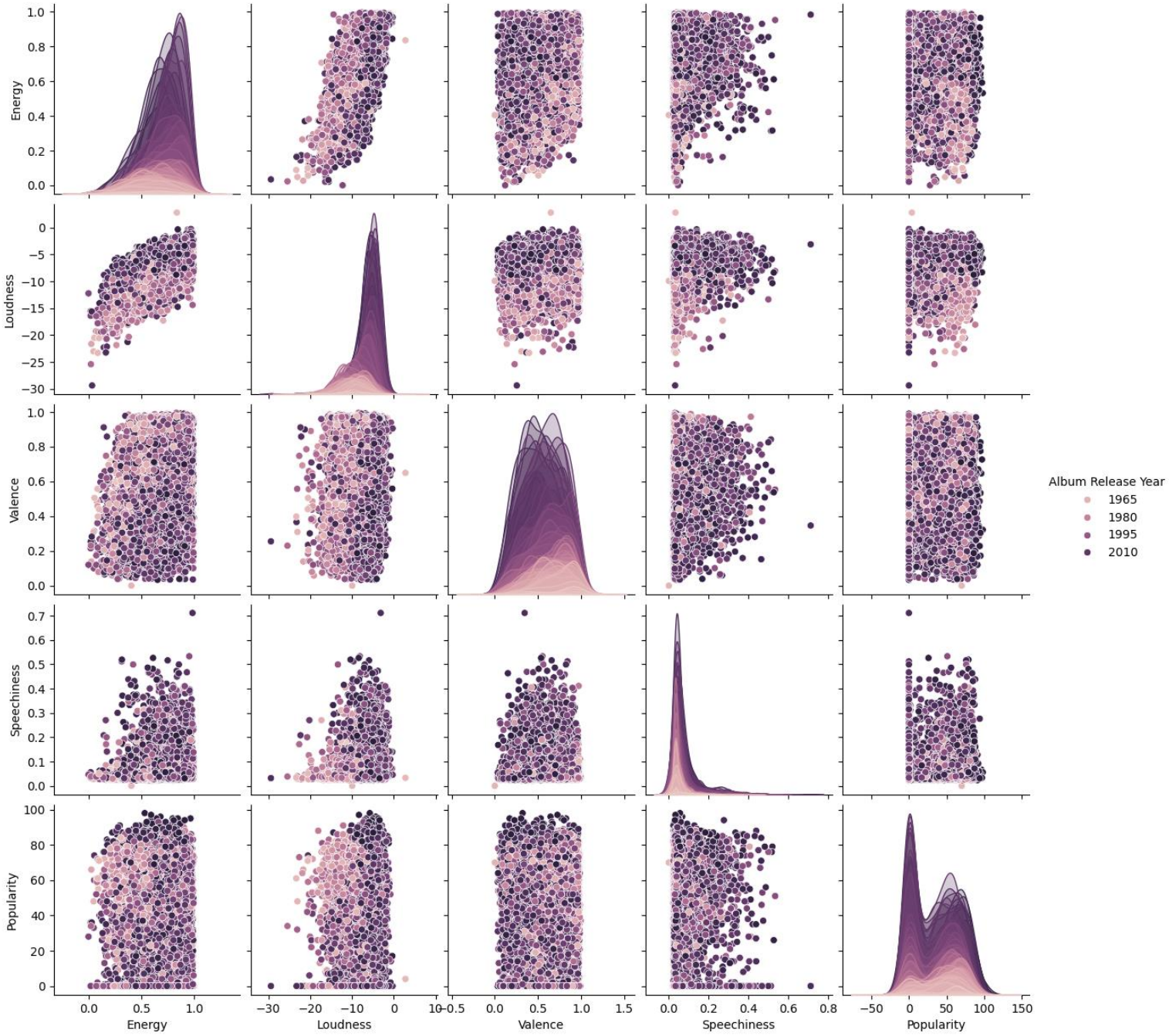

Valence vs Danceability
Danceability
Valence
0.0
0.2
0.4
0.6
0.8
1.0
Explicit
False
True


Album Release Month vs Popularity
Album Release Month
Popularity
August
October
January
December
February
July
September
November
March
June
April
May
0
20
40
60
80
100

Correlation Matrix Heatmap:



Pairwise Scatterplot of: 'Energy', 'Loudness', 'Valence', 'Speechiness', 'Popularity'.



5. Statistical Tests:

For hypothesis testing, a number of statements will be presented and then by utilizing a statistical test, their credibility is evaluated.

Significance Level (**α**) = 0.01

If the calculated p-value is smaller than **α**, the null hypothesis is rejected and the alternative hypothesis is accepted.

5.1 Null Hypothesis 1:

There is no relationship between a song's Energy and its Loudness.

Both Energy and Loudness are quantitative, so an appropriate test would be the Pearson Correlation test.

Test type: **Pearson Correlation**, sample size = **50**

Statistic	P-Value	Result
0.678541	6.1259e-08 < α	Reject Null Hypothesis

Alternative Hypothesis: There is a significant relationship between Energy and Loudness.

5.2 Null Hypothesis 2:

The means of Speechiness are equal for both categories of the Explicit feature.

Speechiness is quantitative and roughly follows a normal distribution. Since Explicitness is a binary categorical feature, the t-test would be appropriate here.

Test type: **T-Test**, sample size = **200**

Statistic	P-Value	Result
2.621459	0.0094 < α	Reject Null Hypothesis

Alternative Hypothesis: There is a significant difference between the means of the two features.

5.3 Null Hypothesis 3:

There is no relationship between a song's Valence and its Danceability.

Both Valence and Danceability are quantitative and they roughly follow a normal distribution, so an appropriate test would be the Pearson Correlation test.

Test type: **Pearson Correlation**, sample size = **50**

Statistic	P-Value	Result
0.558002	2.5504e-05< α	Reject Null Hypothesis

Alternative Hypothesis: There is a significant relationship between Valence and Danceability.

5.4 Null Hypothesis 4:

There is no relationship between a song's Energy and its Acousticness.

Both Energy and Acousticness are quantitative, so an appropriate test would be the Spearman Correlation test.

Test type: **Spearman Correlation**, sample size = **50**

Statistic	P-Value	Result
-0.590668	6.3161e-06 < α	Reject Null Hypothesis

Alternative Hypothesis: There is a significant relationship between Energy and Acousticness.

5.5 Null Hypothesis 5:

There is no relationship between a song's Release Year and its Loudness.

Both the Release Year and Loudness are quantitative, so an appropriate test would be the Spearman Correlation test.

Test type: **Spearman Correlation**, sample size = **50**

Statistic	P-Value	Result
0.507266	0.0001 < α	Reject Null Hypothesis

Alternative Hypothesis: There is a significant relationship between Release Year and Loudness.

6. Conclusion:

Based on the analyzed data, the visualization and the statistical tests, we can conclude that many attributes of the dataset are correlated to each other. Some dependencies are easier to see but others may need deeper analysis.

	Track Name	Artist Name(s)	Album Name	Artist Genres	Track Duration (minutes)	Album Release Month
count	9996	9996	9996	9449	9996	9996
unique	8256	4128	6634	2815	355	12
top	One	Taylor Swift	Greatest Hits	dance pop,pop	03:28	January
freq	9	48	110	254	116	3518

Additionally, many categorical features (like Artist Name) were left relatively unused. There may be relationships between these attributes and the quantitative features that we are unaware of.