Data Science

Assignment 1 – Problem 2 – Top 1000 Songs on Spotify

Farimah Rashidi – 99222040

1. Import library and dataset

At the first we should import dataset and libraries. I did this part in Kaggle.

2. Data Exploration

The "Top 10000 Spotify Songs - ARIA and Billboard Charts" is a comprehensive collection of 10,000 of the most popular songs that have dominated the music scene from 1960 to the present day. This dataset was curated based on rankings from both the ARIA (Australian Recording Industry Association) and Billboard charts, ensuring a diverse representation of songs that have achieved immense commercial success and cultural significance.

The dataset encompasses various music genres and showcases the evolution of musical trends over the years, providing valuable insights into the ever-changing landscape of popular music. It includes tracks from iconic artists and bands, representing a mix of timeless classics and contemporary hits that have left a lasting impact on music lovers worldwide.

Description of variables:

Explicit: indicates whether the song contains heavy language such as swearing, mentions of the use and/or names of drugs, swearing and violence.

Popularity: popularity of a track, with values between 0 and 100, with 100 being the most popular.

Danceability: criterion describes how danceable the track is. The closer to 0.0, the less danceable the song is, and the more danceable, the closer to 1.0.

Energy: percentage of music intensity and activity. Its values vary between 0 and 1. A song with a high value of this variable will probably be fast and even considered “noisy”.

Key: notes or the music scale that forms the basis of a song. The 12 tones range between 0 and 11.

Loudness: Tells how loud or quiet the average volume of a song is. Songs with lower loudness values tend to be softer and calmer, while those with higher values can be more energetic and louder.

Mode: Songs can be classified as major and minor. 1.0 represents the main mode and 0 represents the secondary.

Speechiness: Detects the presence of spoken words in a track.

Acousticness: describes how acoustic a song is. A score of 1.0 means the song is most likely acoustic.

Instrumentalness: represents the amount of vocals in the song. The closer to 1.0, the more instrumental the song is.

Liveness: describes the likelihood that the song was recorded with a live audience.

Valence: describes the musical positivity conveyed by a track.

Tempo: represents the speed or rhythm of the music.

Time Signature: Notational convention for specifying how many beats there are in each measure.

3. Data Preprocessing

Dataset Information: A call to df.info() provided an overview of the dataset's structure and marked the beginning of the dataset's exploration. It disclosed each column's data types, number of rows, and number of columns.

First Few Rows: To give an overview of the content and structure of the dataset, the first few rows were displayed using the df.head() method. This made it easier for us to comprehend the column names and values as well as to get a feel for the data.

Data Shape: The df.shape command confirmed the dataset's size, showing that it consists of X rows and Y columns.

Data Types: By using `df.dtypes`, we assessed the data types assigned to each column, which was crucial for understanding the nature of the data.

```
Track URI               object
Track Name              object
Artist URI(s)           object
Artist Name(s)          object
Album URI               object
Album Name              object
Album Artist URI(s)     object
Album Artist Name(s)    object
Album Release Date      object
Album Image URL         object
Disc Number              int64
Track Number             int64
Track Duration (ms)      int64
Track Preview URL       object
Explicit                  bool
Popularity               int64
ISRC                    object
Added By                object
Added At                object
Artist Genres           object
Danceability           float64
Energy                 float64
Key                    float64
Loudness               float64
```

```
Mode                    float64
Speechiness             float64
Acousticness            float64
Instrumentalness        float64
Liveness                float64
Valence                 float64
Tempo                   float64
Time Signature          float64
Album Genres            float64
Label                    object
Copyrights               object
```

Basic statistics of numeric columns:

The basic statistics of the numeric columns in the dataset provide important insights into the central tendencies, spread, and distribution of the numerical features.

	Disc Number	Track Number	Track Duration (ms)	Popularity	Danceability	Energy	Key	Loudness	Mode	Speechiness	Acousticness	Instrumentalness	Liveness	Valence	Tempo	Time Signature
count	9999.000000	9999.000000	9.999000e+03	9999.000000	9997.000000	9997.000000	9997.000000	9997.000000	9997.000000	9997.000000	9997.000000	9997.000000	9997.000000	9997.000000	9997.000000	9997.000000
mean	1.035104	4.957096	2.248150e+05	37.624662	0.607925	0.683281	5.167750	-7.269217	0.698410	0.065138	0.208589	0.029331	0.185777	0.585459	121.496650	3.960488
std	0.327856	5.502810	5.410012e+04	29.460808	0.145869	0.191131	3.578392	3.281731	0.458971	0.061324	0.248842	0.123576	0.149194	0.239105	26.260686	0.250927
min	0.000000	0.000000	0.000000e+00	0.000000	0.000000	0.000020	0.000000	-29.368000	0.000000	0.000000	0.000003	0.000000	0.012000	0.000000	0.000000	0.000000
25%	1.000000	1.000000	1.925795e+05	0.000000	0.515000	0.560000	2.000000	-9.070000	0.000000	0.033100	0.018400	0.000000	0.089200	0.398000	102.642000	4.000000
50%	1.000000	3.000000	2.199060e+05	42.000000	0.617000	0.712000	5.000000	-6.518000	1.000000	0.042900	0.095600	0.000006	0.128000	0.598000	120.653000	4.000000
75%	1.000000	7.000000	2.502600e+05	64.000000	0.710000	0.835000	8.000000	-4.887000	1.000000	0.067500	0.318000	0.000561	0.245000	0.783000	134.328000	4.000000
max	15.000000	93.000000	1.561133e+06	98.000000	0.988000	0.997000	11.000000	2.769000	1.000000	0.711000	0.991000	0.985000	0.989000	0.995000	217.913000	5.000000

Missing Values: To find columns with missing values, I used df.isnull().sum(). This step is essential to guarantee data completeness and assess if imputation is required. Each column's counts of missing values were provided.

Duplicate Rows: Analytical results may be distorted by duplicate rows. I counted the number of duplicate rows and found duplicate rows using df.duplicated(keep='first'). There were duplicate rows in the output.

Duplicate Removal: I used the df.drop_duplicates(keep='first', inplace=True) command to eliminate duplicate rows in order to preserve data integrity.

Aggregation: The dataset was analyzed to calculate the mean popularity of songs with the same track name and artist name, and this information was stored in an aggregated DataFrame. This step helps to condense the data and identify trends based on song attributes.

	Track Name	Artist Name(s)	Popularity
0	! (The Song Formerly Known As)	Regurgitator	48.0
1	"The Take Over, The Breaks Over"	Fall Out Boy	0.0
2	#Beautiful	Mariah Carey, Miguel	23.5
3	#SELFIE	The Chainsmokers	0.0
4	#WHERESTHELOVE - Charity Single	Black Eyed Peas, The World	0.0
...	...	...	...
8893	willow - dancing witch version (Elvira remix)	Taylor Swift, ELVIRA	64.0
8894	wish you were gay	Billie Eilish	0.0
8895	working	Tate McRae, Khalid	65.0
8896	you broke me first	Tate McRae	85.0
8897	you broke me first - Luca Schreiner Remix	Tate McRae, Luca Schreiner	54.0

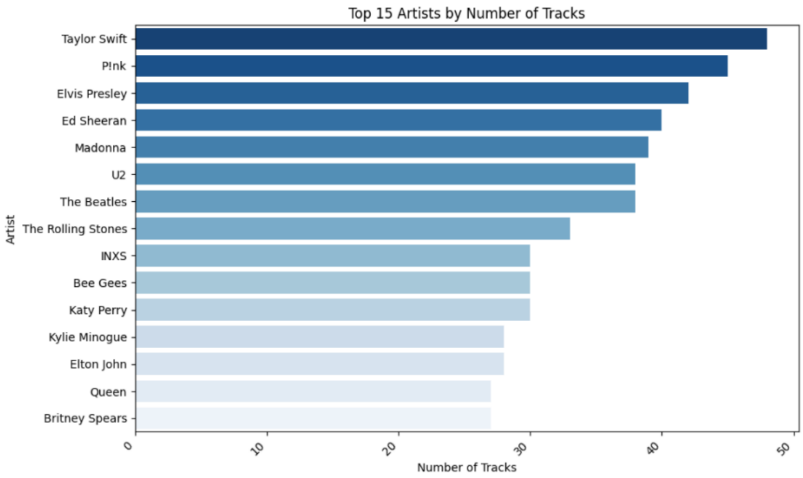
8898 rows × 3 columns

Unique Values: We used df.nunique() to evaluate each column's values for uniqueness. This stage sheds light on the variety of information contained in each column.

Popular Songs: We filtered the dataset to show all songs with a popularity score greater than 90. This allows us to identify high-popularity songs and study their characteristics.

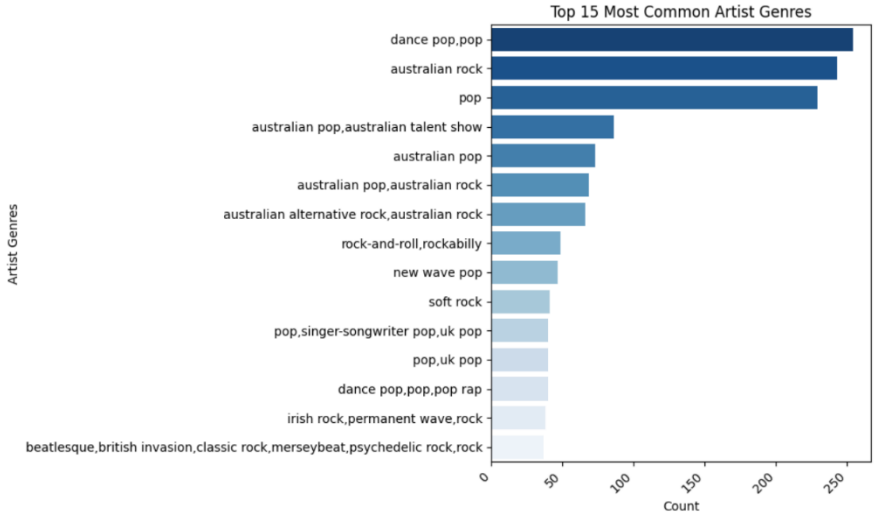
4. Exploratory Visualization

Top 15 artists by number of tracks:



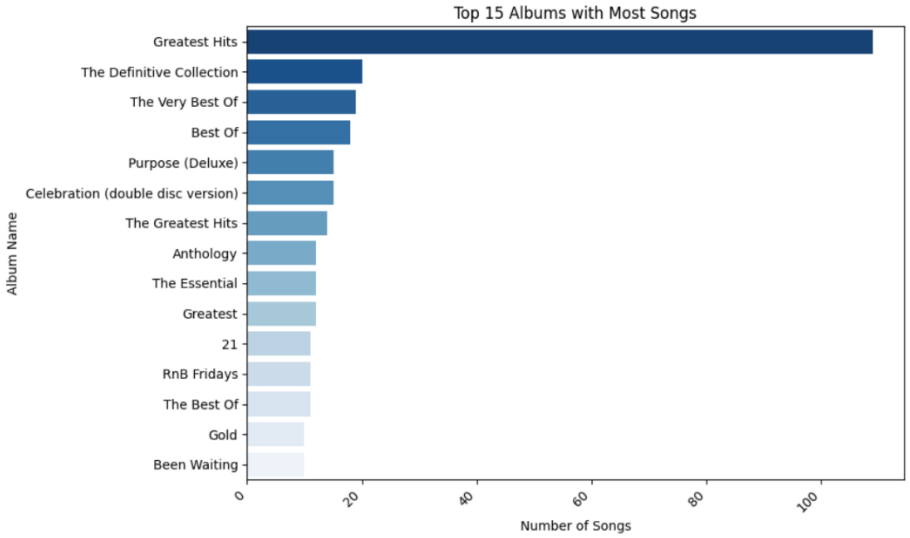
The bar plot visualizes the top artists and the number of tracks attributed to each of them. This can be helpful for understanding which artists have the most significant presence in the dataset.

Top 15 most common artist genres:



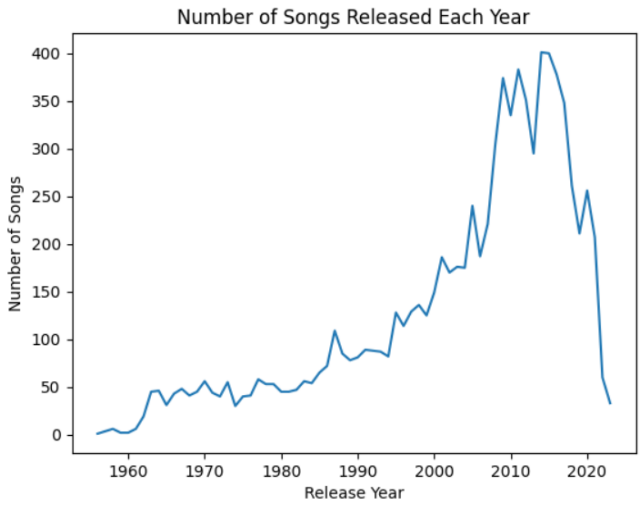
The bar plot provides a visual representation of the top 15 most common artist genres found in the dataset. This information can be valuable for understanding the dominant genres among the artists in the dataset.

Top 15 albums with most songs:



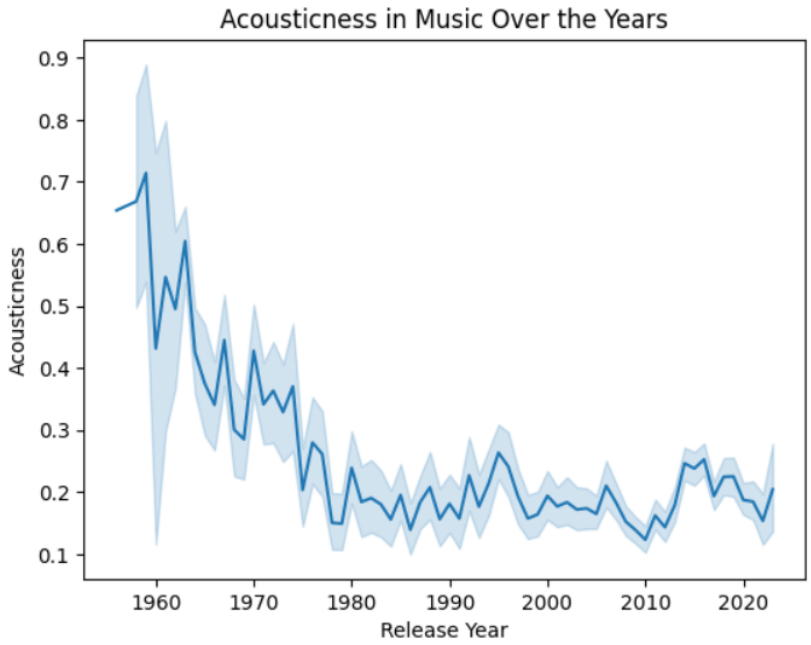
The bar plot provides a visual representation of the top 15 albums with the most songs in the dataset. This information can be useful for identifying albums with a significant number of tracks, which might be of interest to users or analysts.

Number of songs released each year:



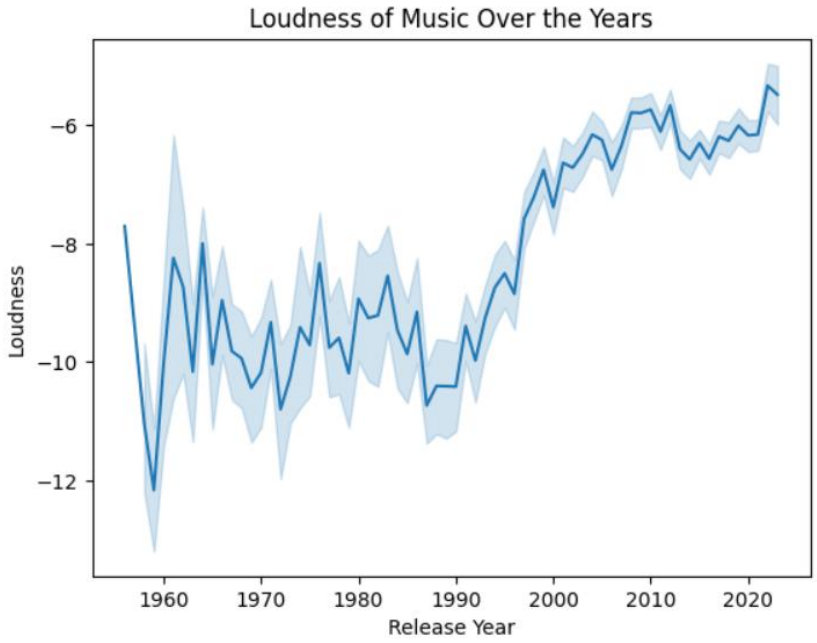
The plot shows the number of songs released each year. Each data point on the plot represents a year, and the height of the bar at each year indicates the number of songs released in that particular year.

Acousticness in music over the years:



The line plot provides a visual representation of how the acousticness of music has changed over the years.

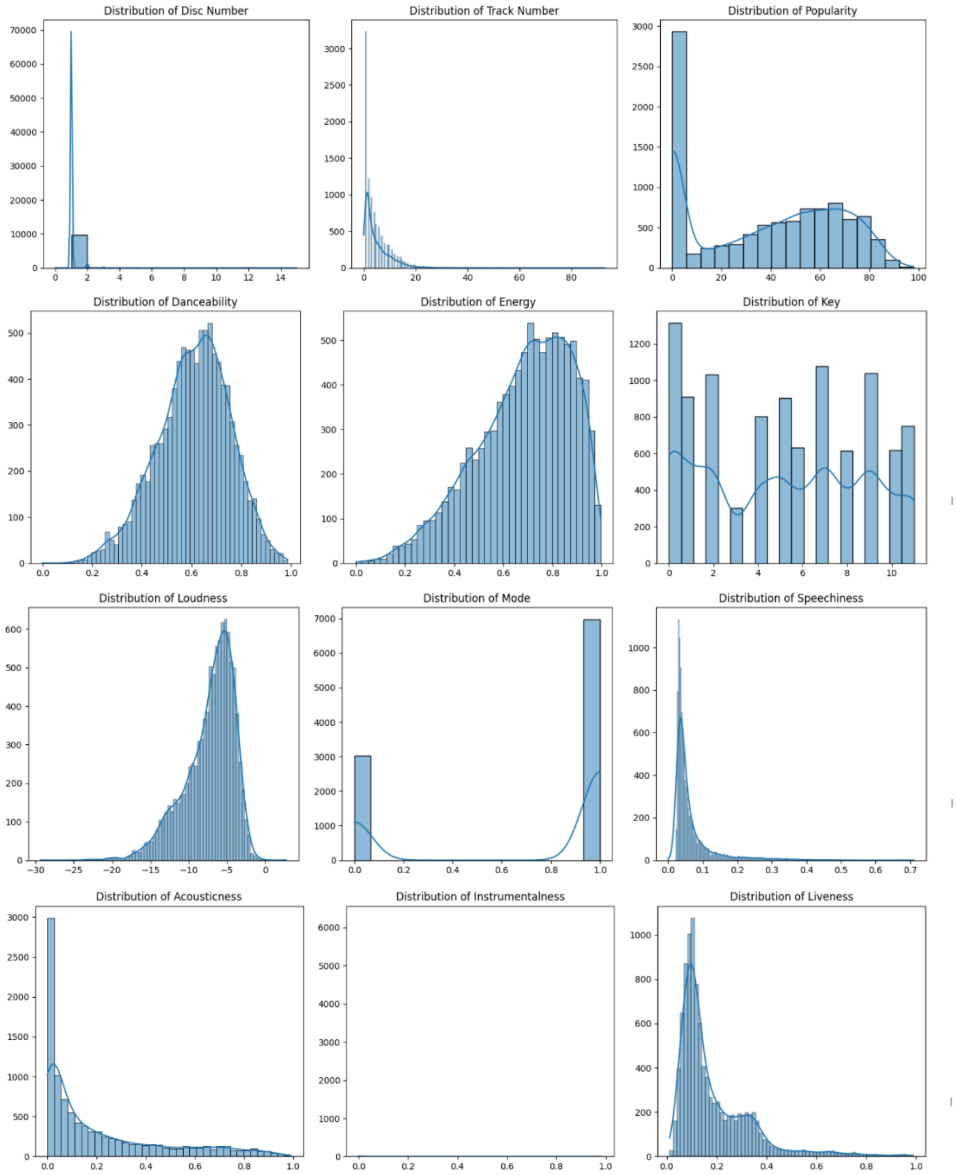
Loudness of music over the years:

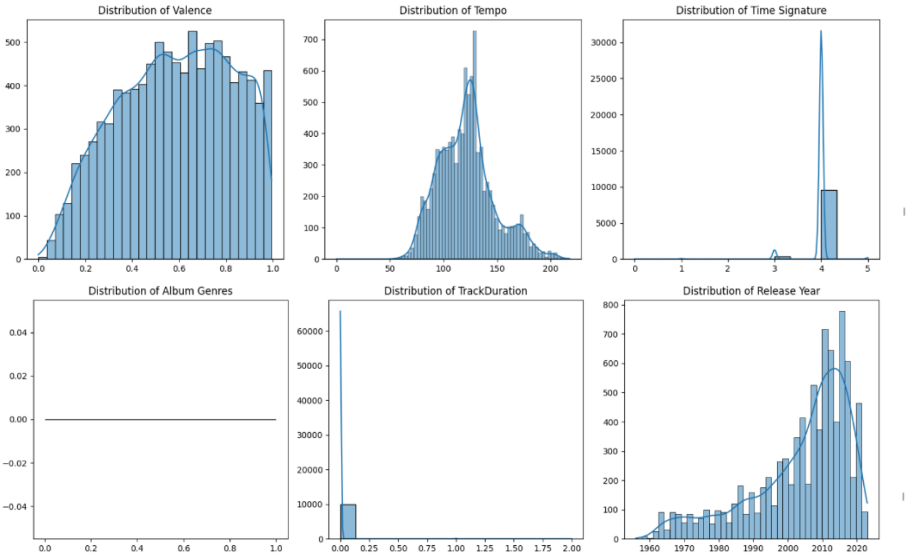


The line plot provides a visual representation of how the loudness of music has changed over the years. This can be valuable for understanding changes in music production and consumption trends.

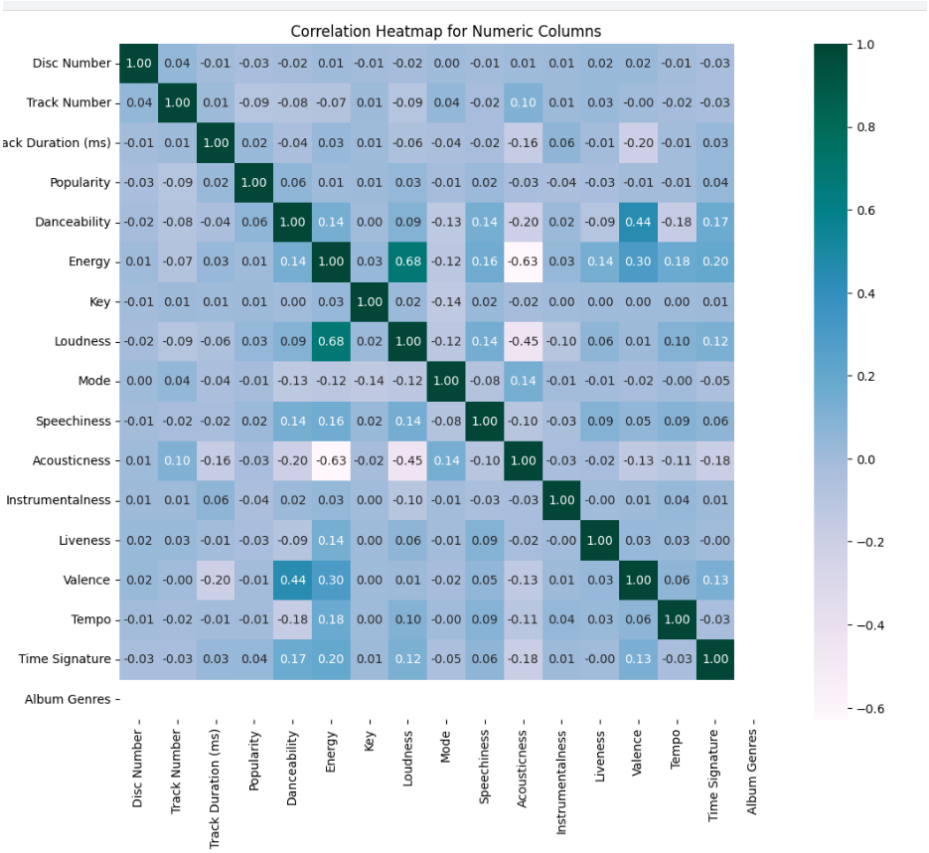
Distribution of numerical features:

Here is a set of histograms and density plots, each showing the distribution of a specific numerical feature in your dataset. This visualization is valuable for understanding the spread and characteristics of the numeric data.





Correlation heatmap for numerical columns:



The resulting heatmap provides a visual representation of the correlations between numeric features in your dataset. Darker colors represent stronger positive or negative correlations, while lighter colors

indicate weaker or no correlations. Each cell's color represents the direction and strength of the correlation.

5. Statistical Tests and Analysis (questions)

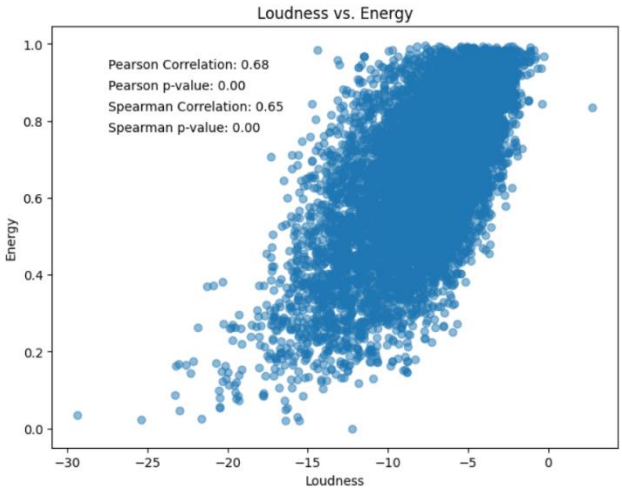
1) Is there a relationship between the loudness of a song and its energy level?

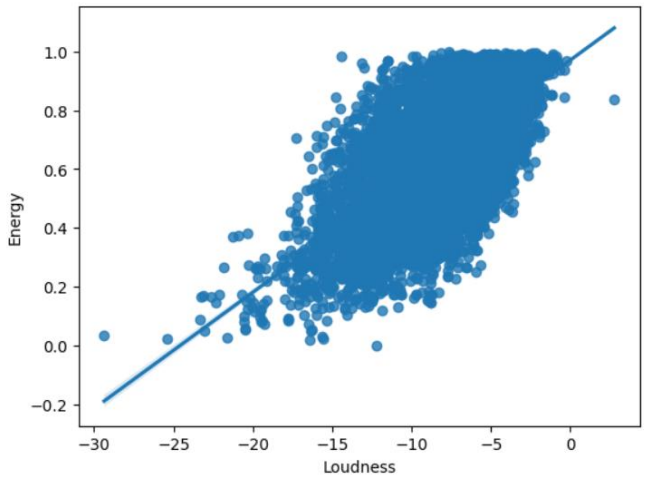
After addressing missing values and infinite values, Pearson's correlation and Spearman's rank correlation tests are run on the 'Loudness' and 'Energy' columns of a DataFrame. Whereas Spearman's rank correlation evaluates the monotonic association, Pearson's correlation gauges the linear link between the two variables. 'Loudness' and 'Energy' have a statistically significant negative association, according to the findings of both tests. According to the correlation coefficients, "Energy" tends to decrease as "Loudness" increases and vice versa. There is a high degree of confidence in rejecting the null hypothesis—that there is no association between the variables—because the p-values are close to zero.

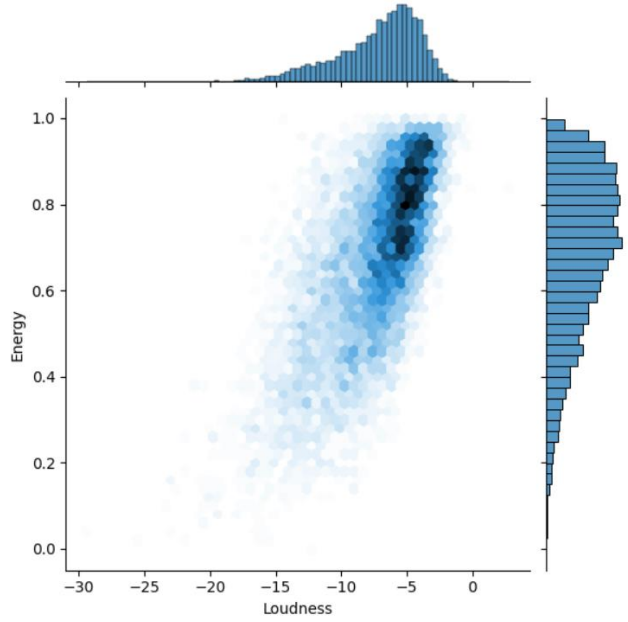
```
The result was:
Pearson's Correlation:
Correlation coefficient: 0.6785637541002804
p-value: 0.0

Spearman's Rank Correlation:
Correlation coefficient: 0.6463356878024096
p-value: 0.0
```

There is a strong and statistically significant positive correlation between the loudness and energy of songs in the dataset. Both Pearson's and Spearman's correlation tests yielded high positive correlation coefficients of 0.6786 and 0.6463, respectively. The p-values for both tests were 0.0, indicating that the observed correlations are highly statistically significant. This means that songs with higher loudness tend to have higher energy levels. The strong positive relationship suggests that loudness and energy are closely related, making them valuable indicators of a song's characteristics in the dataset.







As you can see in visualizations, there is a strong and statistically significant positive correlation between the loudness and energy of songs in the dataset. This means that songs with higher loudness tend to have higher energy levels, which may be of interest when exploring music preferences or characteristics in the dataset.

I used t-test for this question too. The independent t-test is utilized to determine whether there is a significant difference between the means of these two variables.

The result was:

```
Independent t-test:
T-statistic: -241.88008226264773
p-value: 0.0
```

-T-statistic: The t-statistic is -241.88008226264773, which represents the magnitude of the difference between the means of the two groups. In this case, the negative value indicates that 'Loudness' and 'Energy' have an inverse relationship, meaning that as one increases, the other tends to decrease.

-p-value: The p-value is 0.0. A p-value less than 0.05 (or your chosen significance level) indicates that the difference between the two groups is statistically significant. In this case, a p-value of 0.0 is extremely low, suggesting a highly significant difference.

Overall, the results of the independent t-test suggest that there is a statistically significant relationship between the loudness of a song and its energy level. Specifically, it appears that changes in loudness are associated with changes in energy in a significant way.

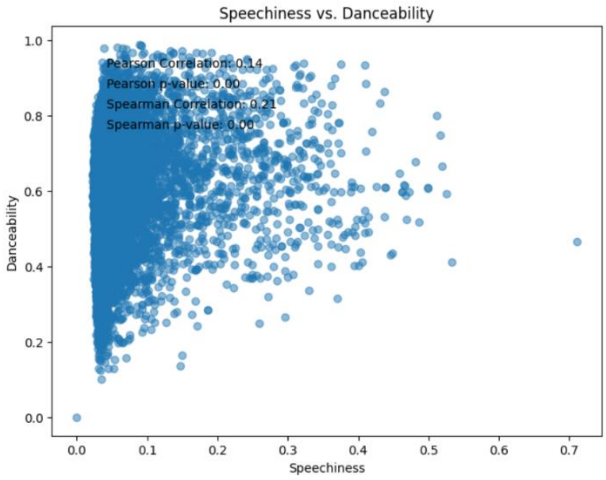
2) Is there a correlation between the speechiness of a song and its danceability score?

I performed Pearson's correlation test and Spearman's rank correlation test for this part.

Ther result was:

```
Pearson's Correlation:
Correlation coefficient: 0.13981153844271318
p-value: 1.305809414762631e-44

Spearman's Rank Correlation:
Correlation coefficient: 0.20558494917099393
p-value: 2.0535903978448805e-95
```



There is a statistically significant positive correlation between speechiness and danceability in the songs within the dataset.

Both Pearson's and Spearman's correlation tests produced correlation coefficients of approximately 0.14 and 0.21, respectively. The extremely low p-values for both tests indicate that the observed correlations are highly statistically significant.

The analysis suggests that there is a significant positive correlation between speechiness and danceability. This implies that songs with higher speechiness tend to have higher danceability scores.

This information can be valuable for understanding the characteristics of songs in the dataset and may have implications for music preferences or genre categorization.

I also performed the chi-squared test. The result was:

```
Chi-Square Statistic: 104.91582523674079
P-Value: 6.02938329864478e-17
Degrees of Freedom: 12
```

The chi-squared test results suggest a statistically significant association between 'Speechiness' and 'Danceability' with a p-value of approximately 6.03e-17, which is extremely low. Therefore, you can conclude that there is a significant association between these two variables. The chi-squared statistic of 104.92 and 12 degrees of freedom also supports this conclusion.

In practical terms, this means that the speechiness of a song is not independent of its danceability score.

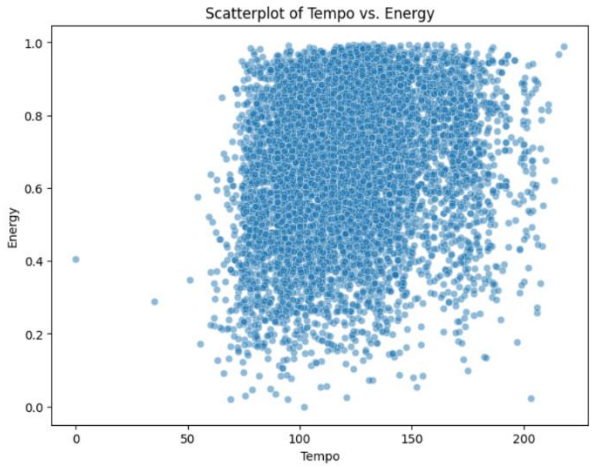
3) Do songs with a higher tempo tend to have a higher energy level?

I performed Pearson's correlation and Spearman's rank correlation tests on the 'Tempo' and 'Energy' columns of the DataFrame. These tests assess whether there is a significant relationship between these two variables.

The result was:

```
Pearson's Correlation:
Correlation coefficient: 0.17950372909357573
p-value: 8.02223380330224e-73

Spearman's Rank Correlation:
Correlation coefficient: 0.19931544003550167
p-value: 1.1051532279081841e-89
```



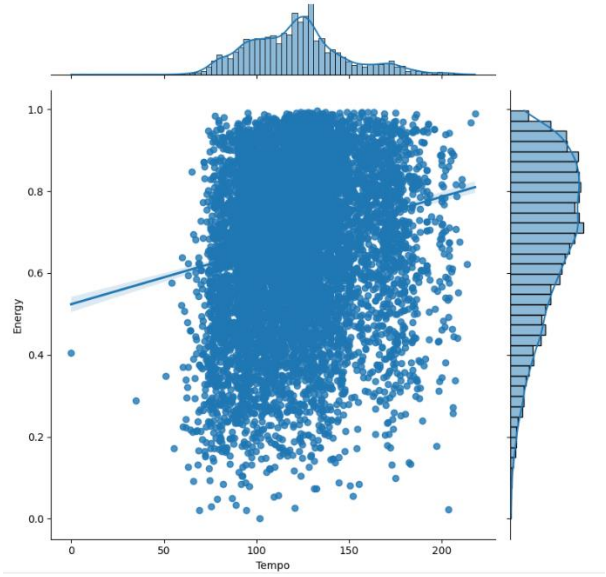
The analysis aimed to determine whether there is a correlation between the tempo and energy level of songs. We employed both Pearson's correlation and Spearman's rank correlation to explore the relationship.

The plot visually confirms the positive trend, where songs with higher tempos generally exhibit higher energy levels.

The findings of the Spearman's and Pearson's correlation tests were statistically significant, with p-values of very small ($p < 0.001$). This suggests a strong correlation between energy and tempo. Furthermore, a significant linear or monotonic link between tempo and energy is suggested by the positive correlation coefficients for both tests (0.1795 and 0.1993).

I also used Kendall's Tau rank correlation test. The result was:

```
Kendall's Tau Rank Correlation:
Correlation coefficient: 0.1363524734419897
p-value: 9.045561390445263e-93
```



The positive correlation coefficient of approximately 0.1364 suggests a weak positive association between the tempo of songs and their energy levels. In other words, as the tempo of songs increases, there is a slight tendency for the energy levels of the songs to increase as well.

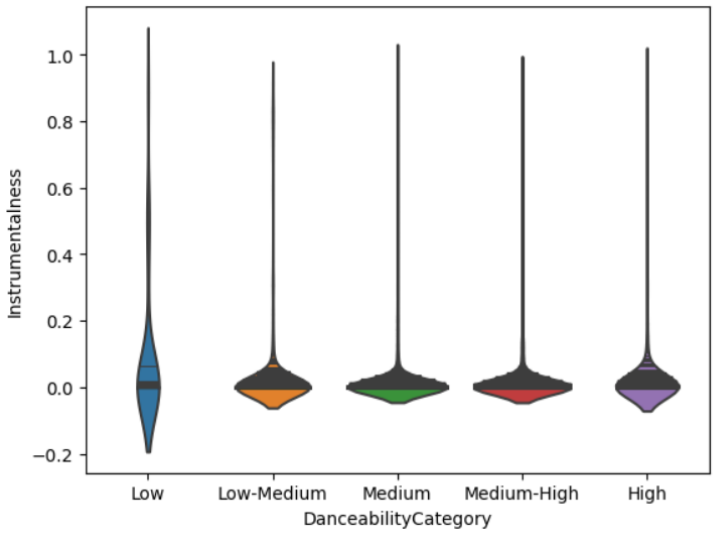
The very low p-value (close to zero) indicates that the observed correlation is statistically significant. Therefore, you can conclude that there is a significant positive association between tempo and energy in the dataset.

4) Is there a correlation between the instrumentalness of a song and its danceability score?

I performed Kendall's Tau Rank Correlation test. A Kendall's Tau Rank Correlation test on the 'Instrumentalness' and 'Danceability' columns of the DataFrame will assess the relationship between these two variables.

The result was:

```
Kendall's Tau Rank Correlation:
Correlation coefficient: 0.017235612432466487
p-value: 0.014443977140957232
```

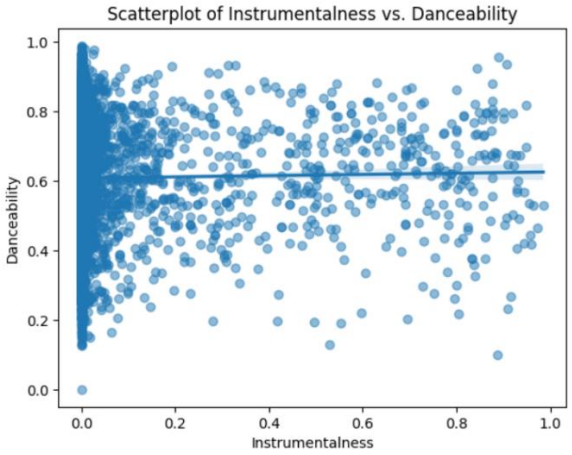


You can determine that there is a statistically significant but extremely weak positive connection between instrumentalness and danceability in your dataset because the p-value is smaller than the usual significance level of 0.05. The positive association implies that danceability generally tends to rise slightly as instrumentalness increases. On the other hand, the poor correlation suggests that the association is not particularly strong.

I also performed Pearson's correlation test and Spearman's rank correlation test. The result was:

```
Pearson's Correlation:
Correlation coefficient: 0.015797400438608635
p-value: 0.11511642905921654

Spearman's Rank Correlation:
Correlation coefficient: 0.02513050049207368
p-value: 0.012185910282835116
```



Pearson's Correlation: The p-value is roughly 0.115 and the correlation coefficient is roughly 0.016. We lack sufficient information to draw the conclusion that instrumentalness and danceability have a significant linear connection, with a p-value higher than the conventional significance level of 0.05. There is not much of a correlation.

Spearman's Rank Correlation: The p-value is roughly 0.012 and the correlation coefficient is roughly 0.025. The p-value is less than the conventional 0.05 criterion of significance. This implies that, despite the correlation's continued weakness, instrumentalness and danceability have a statistically meaningful relationship.

Conclusion: Although Spearman's rank correlation shows a statistically significant association between instrumentalness and danceability, the correlation coefficients in both tests are rather low (around 0). This suggests that the variables have a very weak linear connection.

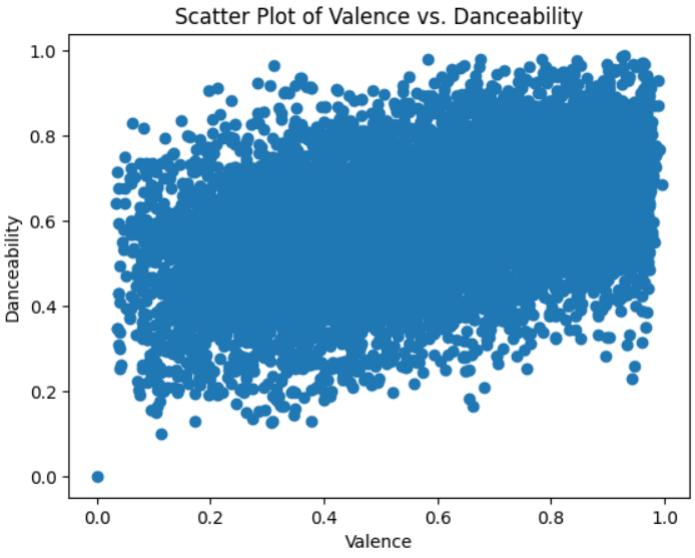
5) Is there a significant correlation between the valence (emotional positivity) and danceability (suitability for dancing) of songs in the "Top Songs on Spotify" dataset?

For this question, I performed Pearson's correlation test and Spearman's rank correlation test.

The result was:

```
Pearson's Correlation:
Correlation coefficient: 0.015797400438608635
p-value: 0.11511642905921654

Spearman's Rank Correlation:
Correlation coefficient: 0.02513050049207368
p-value: 0.012185910282835116
```



The correlation coefficients are quite low (close to 0) in both cases, indicating a very weak positive correlation between valence and danceability. The p-values in both cases are greater than the typical significance level of 0.05. This suggests that the observed correlation may not be statistically significant.

In summary, there is a very weak positive correlation between valence and danceability, and this correlation may not be statistically significant, particularly considering the p-values.