Data Analytics on Spotify using Python

Aditya Chawla

Graduate StudentA , Master of Science in Business Analytics

GBA 6070: Programming Foundation in Business QAnalytics

Dr. Fadi Batarseh

November 26, 2022

Introduction -

Music has been a form of entertainment for ages for human beings. These days there are various music platforms like youtube, apple music, pandora and Spotify that are popular among the general population. A recent study has shown that the music industry has grown by 18.5 percent in the last one year ("Global music report", n.d.) which sparked my research interest into the music industry. The revenue growth of the music industry is at staggering $25.9 billion as compared to last year's $21.9 billion. . Given all the revenue growth among the music giants, I decided to work on analyzing the data from Spotify; it offers digital music, podcast and video services.

The dataset of Spotify is extracted from Kaggle.

Business Framing-

This project will focus on two major questions-

1. A. Are songs with more energy more danceable? I will explore this question because I believe songs with more energy and are more danceable.

   B. Are songs with more loudness more danceable? I believe that if songs have more loudness, one can dance to them and that is why it has been explored in this report.

2. What is the relationship between danceability, energy, loudness, duration of song, explicitly and popularity of music? Using a heat map, we will illustrate whether these elements of music have a positive or a negative relationship with each other. I am interested in exploring these questions as I want to learn how this creative industry of music works and what components make a great song in all aspects?

3. Are songs with more explicit words more popular than those which do not have explicit words? I want to explore this idea because a decade back in India, such songs were more popular among the youth.

I also want to explore the relationship between duration of song and popularity? Are songs with more or less duration more popular? Are songs with more explicit words longer than songs with less explicitly? As per my observation, the duration of songs have seen a downward trend meaning generally new songs are shorter than old songs but they are more popular than old songs. Hence, I wanted to explore this idea. We have seen that nowadays, more explicit songs are made these days. I wanted to learn if they are shorter or longer than songs that do not have more explicit words.

Understanding data-

Kaggle offers free online data and the website is used by data scientists and machine learning enthusiasts. We are using the Spotify dataset which comes from Kaggle. It has been collected from Spotify and uploaded on Kaggle. The dataset has 21 variables which are described below-

- Unnamed- These represent the number of tracks which the dataset has. This is an int object.
- track_id- This is an object which displays different track ids assigned to the tracks. The track ids are the spotify id for the track. Track id is an object.
- Artists- These are the names of the artists who have performed the song or songs. Some songs have multiple artists that have been separated by a semicolon (;). Artists is also an object type.
- Album name- The name of the album that contains the song. This is also an object type of data.

- Popularity- The value of popularity is between 0 and 100, with 0 being the least popular and 100 being the most popular. An algorithm is used to calculate the popularity. The song which has been played the most has the popularity rating of 100 whereas the song which has been played the least has the lowest ranking. If there is a repeated track, for example if it is played by a single artist and is in an album, they are rated independently. Track popularity shows the popularity of artists and album popularity. Popularity is an integer value, int64 data type.
- Duration_ms- This is the length of the track in milliseconds. Duration is in numbers and hence this is an integer type data (int64).
- Explicit- This is shown by a boolean i.e., true or false. Explicit represents whether a song has explicit words or not and these are shown by true or false. This data type is a boolean as it is represented by True or False.
- danceability- danceability means how much a song is danceable. Musical components such as beat strength, rhythm stability, tempo and overall regularity describes danceability. Highest danceable songs are shown by the value of 1.0 and on the other hand, lost danceable songs are shown by value of 0.0. This data type is also int64 type, which means an integer. This is a float type of dataset as values are in decimals.
- energy: Energy is a measure from 0.0 to 1.0 and represents a perceptual measure of intensity and activity. Typically, energetic tracks feel fast, loud, and noisy. For example, death metal has high energy, while a Bach prelude scores low on the scale. This is a float type of dataset as values are in decimals. This is an integer type of dataset as it is in numbers.
- key: The key the track is in. Integers map to pitches using standard Pitch Class notation. E.g. `0 = C`, `1 = C♯/D♭`, `2 = D`, and so on. If no key was detected, the value is -1. This is an integer type of dataset as it is in positive or negative numbers.

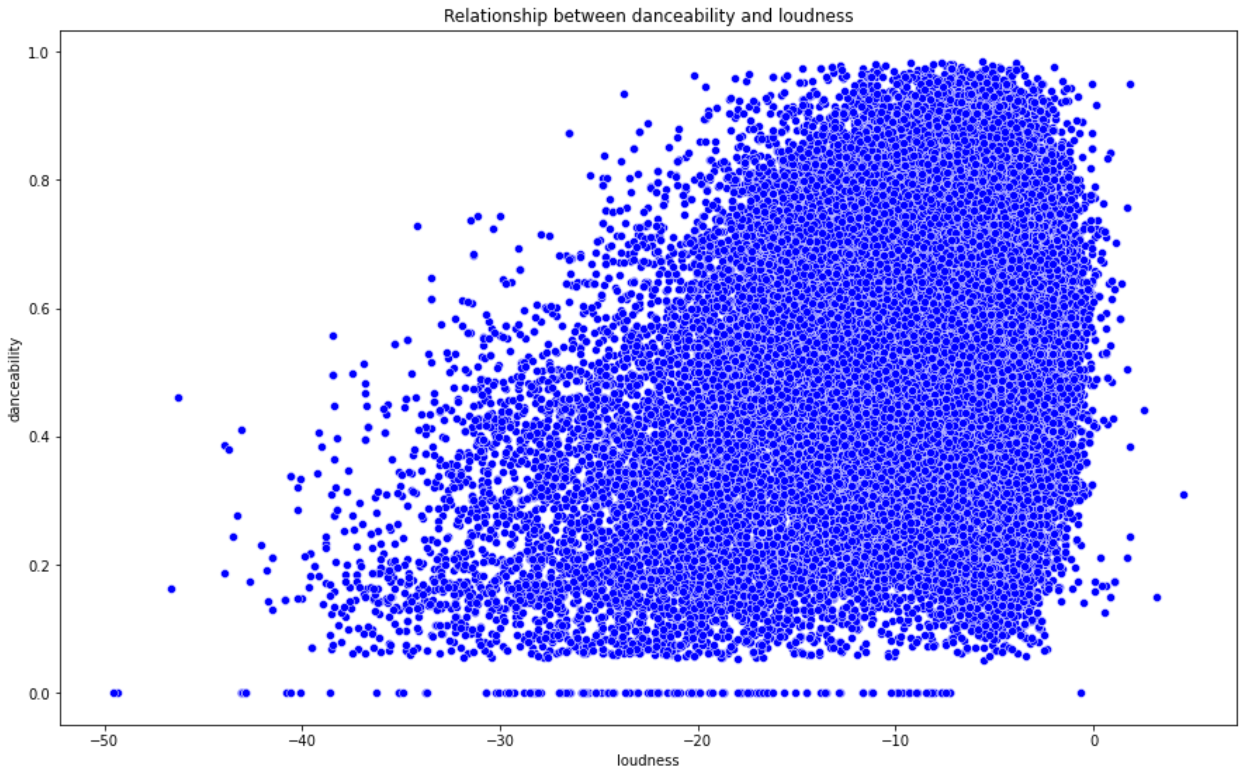
- loudness: The overall loudness of a track in decibels (dB). This is a float type of dataset.
- mode: Mode indicates the modality (major or minor) of a track, the type of scale from which its melodic content is derived. Major is represented by 1 and minor is 0. This is an int type of dataset.
- speechiness: Speechiness detects the presence of spoken words in a track. The more exclusively speech-like the recording (e.g. talk show, audio book, poetry), the closer to 1.0 the attribute value. Values above 0.66 describe tracks that are probably made entirely of spoken words. Values between 0.33 and 0.66 describe tracks that may contain both music and speech, either in sections or layered, including such cases as rap music. Values below 0.33 most likely represent music and other non-speech-like tracks. This is a float type of dataset.
- acousticness: A confidence measure from 0.0 to 1.0 of whether the track is acoustic. 1.0 represents high confidence the track is acoustic. This is a float type of dataset.
- instrumentalness: Predicts whether a track contains no vocals. "Ooh" and "aah" sounds are treated as instrumental in this context. Rap or spoken word tracks are clearly "vocal". The closer the instrumentalness value is to 1.0, the greater likelihood the track contains no vocal content. This is a float type of dataset.
- liveness: Detects the presence of an audience in the recording. Higher liveness values represent an increased probability that the track was performed live. A value above 0.8 provides strong likelihood that the track is live. This is a float type of dataset.
- valence: A measure from 0.0 to 1.0 describing the musical positiveness conveyed by a track. Tracks with high valence sound more positive (e.g. happy, cheerful, euphoric), while tracks with low valence sound more negative (e.g. sad, depressed, angry. This is a float type of dataset.

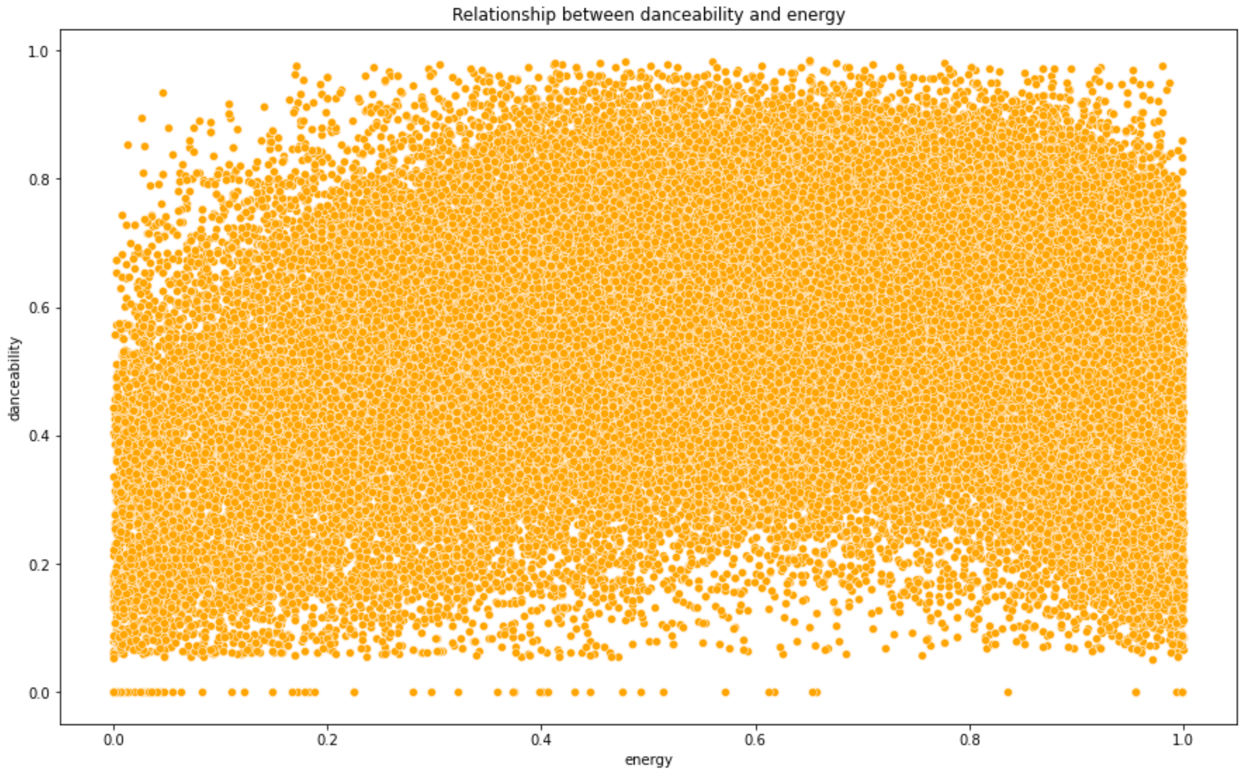
- tempo: The overall estimated tempo of a track in beats per minute (BPM). In musical terminology, tempo is the speed or pace of a given piece and derives directly from the average beat duration. This is a float type of dataset.
- time_signature: An estimated time signature. The time signature (meter) is a notational convention to specify how many beats are in each bar (or measure). The time signature ranges from 3 to 7 indicating time signatures of `3/4`, to `7/4`. This is an int type of dataset.
- track_genre: The genre in which the track belongs. This is an object type of dataset.

The dataset consists of 114 different types of genres of songs. The data of some columns have been dropped such as track id, key, unnamed (used for track numbers), mode, speechiness, acousticness, instrumentalness, liveness, valence, tempo, and time signature to have a better understanding of data. We have used is.na to look at missing values and dtypes to know the types of data in the dataset. The dataset has 114 rows (instances) and 21 columns.

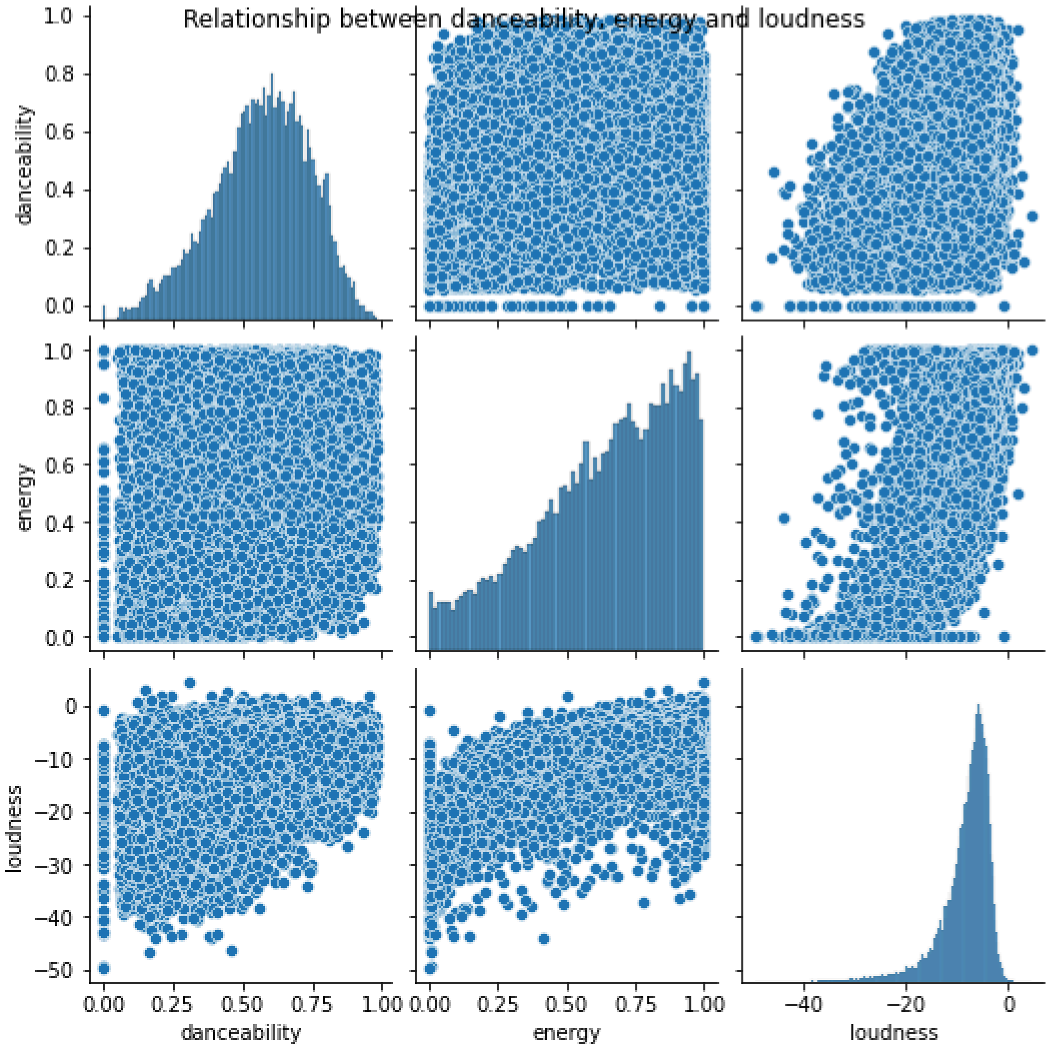
Findings of data-

Firstly, to learn the relationship between danceability, energy and loudness, we looked at the mean, mode and standard deviation of these variables. As per the dataset, the danceability of songs have a mean of 0.57 which means that most of the songs are reasonably good for dancing. The mean for energy is 0.64 which suggests that most songs in the data also have good energy. Mean of loudness is -8.26. This dataset has many outliers and that is why a good finding cannot be generated. Mode of danceability is 0.647 which means it is the most repeated number. For energy, it is 0.876 and for loudness, it is -5.662 which shows the most value in the dataset. Danceability has a standard deviation of 0.173542 which means that most songs are clustered around mean in terms of danceability. Energy also has a relatively small standard deviation (0.251529) which depicts that most songs have an energy level close to mean. As there are many outliers and the range of values of standard deviation of loudness is big, we cannot derive any positive or negative insights from standard deviation of loudness.

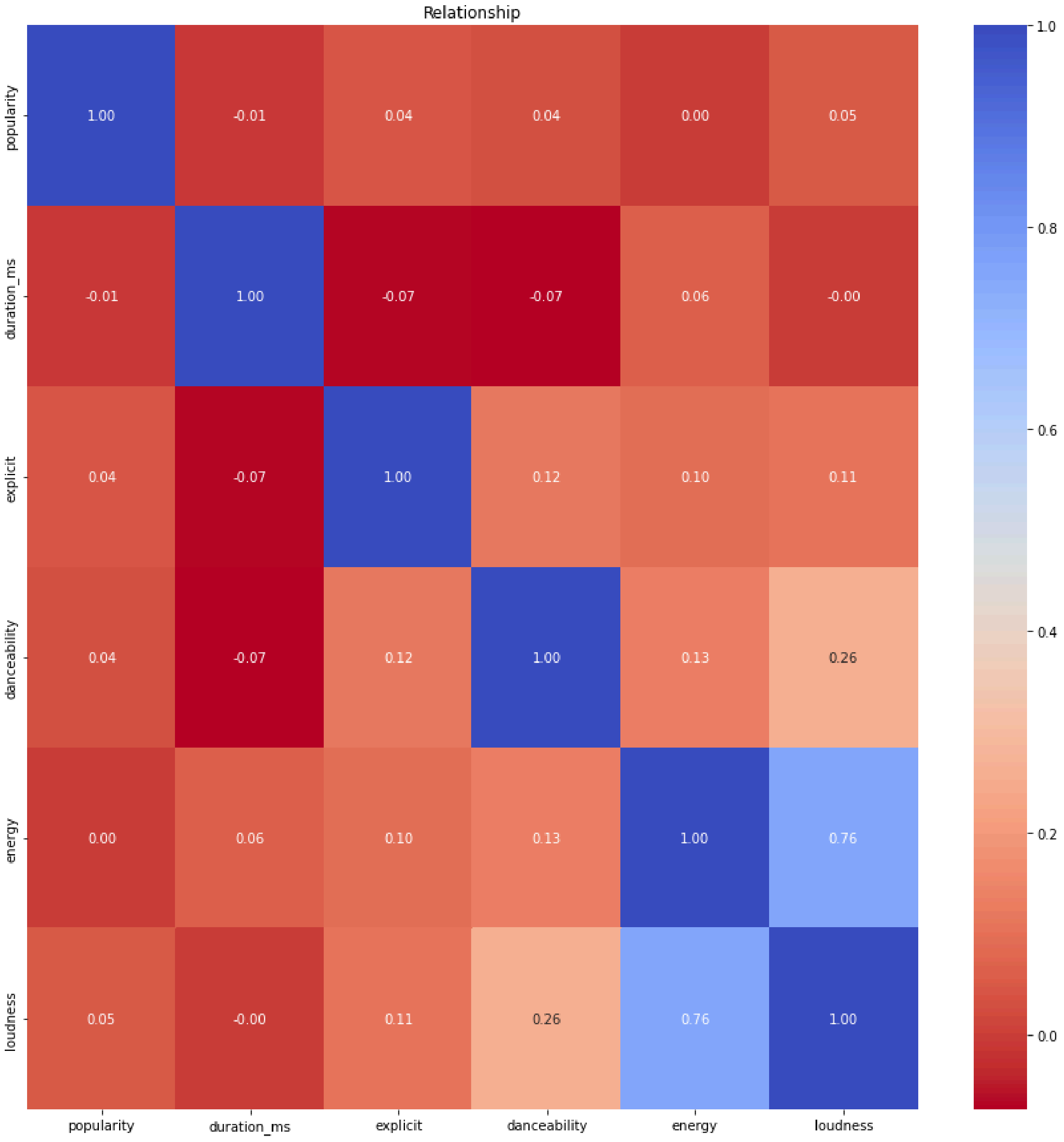

Relationship between danceability and loudness
danceability
1.0
0.8
0.6
0.4
0.2
0.0
−50
−40
−30
−20
−10
0
loudness


Relationship between danceability and energy
danceability
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0
energy

As we can see in figure one (with blue dots), as the loudness increases, the danceability increases. However, there are too many outliers and thus we cannot generalize this statement. In the second graph above, at the bottom, we can see that there are too many outliers. Some songs with maximum energy are not danceable. The whole data is clustered and nothing can be derived from the scatter plot above.

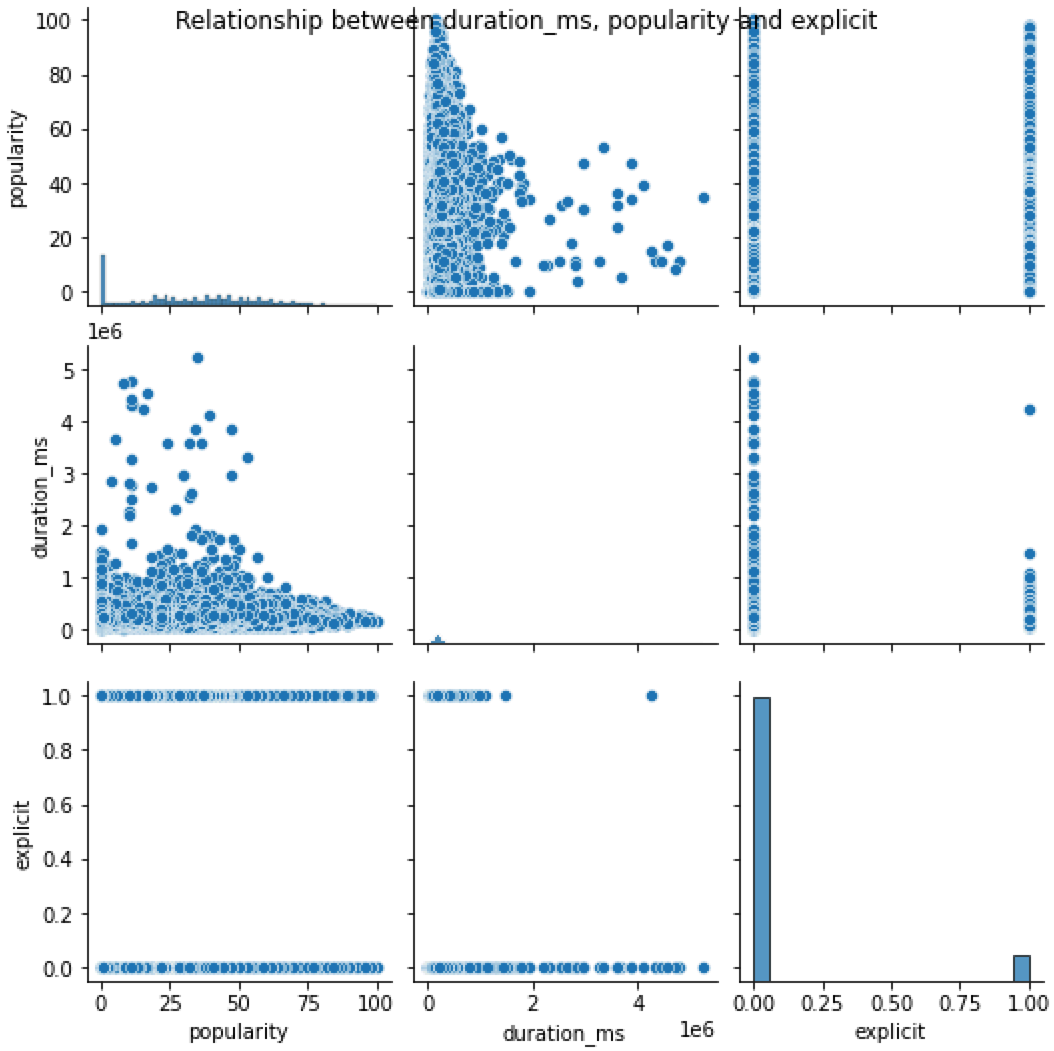


We have also used the pair plot to do visualizations. As per the second graph in row 1, songs with low and high energy both are danceable. It cannot be said whether songs with more energy are danceable or not. As per the row 2 third graph, as loudness increases energy also increases. However, there are some outliers in the dataset. As per row 3 graph 1, as loudness increases, danceability decreases.



A heat map can be defined as a graphical representation of data based on user behavior. Heat maps utilize a system of color-coding to represent the user behaviors – red means that the user activity is high and blue means user activity is low. We have used a heat map to derive relationships between various elements such as popularity, duration, explicit, danceability, energy and loudness. As per the heat map, energy and loudness does not have a strong positive relationship as the value is 0.76. There is a strong positive relationship between duration and popularity. The loudness and danceability have a score of 0.26 which means that there is a moderate relationship between them as per the heat map. It is found that songs that are explicit have more loudness and its score is 0.11. Danceability and energy have a strong positive relationship and their score is 0.13. Duration and danceability score is -0.07 which means they have a negative relationship. Explicit and duration of songs also have a negative relationship as the score is below 0 i.e., -0.07. Energy has a weak but positive relationship with duration with a score of 0.06. Danceability has a weak but positive relationship with popularity with a score of 0.04. Explicit songs have a weak but positive relationship with popularity with a score of 0.04. Duration and popularity have a negative relationship with a score of -0.01. Loudness of songs have a weak but positive relationship with popularity with a score of 0.05. Energy and popularity has zero relationship as the value is 0.0. Duration and loudness have a score of 0.0 which means that there is no relationship. Explicit songs with danceability have a score of 0.12 which means that they have a reasonably good positive relationship. Energy and explicit have a score of 0.10 which means that they also have a positive relationship.

In the third question, we find that the mean of explicit in songs is 0.08 which is low. It means that most songs do not have explicit words. The mean duration of songs is 228029.15, which is approximately 3.80 minutes. Popularity rating's mean is 33.23 which means that most songs in the dataset have low popularity as maximum popularity can be 100. Mode for explicit is false which means that most songs do not have explicit words in them. The mode for duration is 162897 which means that this is the most repeated value and represents the duration of song in milliseconds. Mode of popularity is 0 which means that many songs in the dataset have zero popularity. Standard of explicit is 0.28 approximately which means that there is moderate data around mean of explicit songs. Standard deviation of duration_ms (duration of song in milliseconds) is 108 seconds. This is because of outliers. As there are many outliers which have either very small or big duration. Even the standard deviation of popularity is too high due to outliers and nothing useful can be derived from data.



If we look at the third graph in row one, it suggests that songs with no explicit words are also popular like songs with explicit songs. Moreover, we can generalize as the dataset might be missing too many songs with explicit words. Graph 1 in row 2 indicates that as the duration of songs increases, the popularity decreases and vice versa but the dataset has many outliers. Graph 2 in row three indicates that songs that have explicit words are usually shorter than songs with no explicit words. There are only a few outliers. Therefore, we can generalize that.

Conclusion-

Firstly, we imported the libraries such as pandas, numpy, matplotlib, seaborn and warnings. After importing libraries, we uploaded the dataset. Post that, we looked at instances using shape. Then, we looked at the dataset using head. Next step was to find missing values using the is.na function. Then, we used the dtypes to find the type of dataset. Then, I used to drop multiple columns to clean the dataset. Then, we glanced at the dataset which had the columns after cleaning up. After that, we find the mean, mode and standard deviation of loudness, energy and danceability, and explicitly, duration and popularity. Then, we used .describe to look at the other statistical variables such as count, min, max and quartiles. This helped me in finding how the dataset is spread out, what are most repeated values and their mean which was important because it helps in understanding the distribution.We used unique and len functions to find out the number of genres types of genres. Then, we removed the track id as it was irrelevant. I used the drop function and data became clean for analysis. Then, we imported packages such as seaborn, matplotlib.pyplot as plt and warnings for visualization. Warnings were necessary so that visualization could be better and there were no errors. Then, we used sns.pairplot to visualize the relationship between danceability, energy and loudness. We used scatter plots to illustrate the relationship between danceability and energy and danceability and loudness. Then, I imported more libraries such as %matplotlib inline, numpy, os, seaborn and matplotlib.pyplot. This was important to see how much activity was there between the 6 variables- danceability, energy, loudness, duration, explicit and popularity. Then, I converted the explicit data values from True and false to 0 and 1 using astype. This helped me in getting better results for visualizations. Then, we created a pair plot to see the relationship between duration, explicit and popularity.

This helped me derive insights about the data and come up with analysis. This was the last part of the project. Findings have been summarized under findings separately.

We suggest that data of a particular decade can be taken which has a popularity of at least 32 as that was the mean for this data so that there are fewer outliers. Also, this dataset did not have songs from other countries which means the results for other country songs can differ.

Reflection-

I learned that a pair plot could be a very useful and powerful tool even though outliers are there. I learned how to represent a dataset using statistical tools and visualizations. I have changed the value from False and True to 0 and 1 in the explicit columns to find out the relationship between explicitly, duration and popularity of songs which was the most interesting part for me.

There are few drawbacks to the data. It does not include all the artists and their songs in the world. This was an interesting project and it helped me in improving my coding skills and I now have a better understanding of how to clean data, visualize it, remove irrelevant data and convert boolean data into integer type. I enjoyed working on this project.

References

MaharshiPandya. (2022, October 22). *spotify tracks dataset*. Kaggle. Retrieved November 16, 2022, from https://www.kaggle.com/datasets/maharshipandya/-spotify-tracks-dataset

Richter, F. (2022, April 8). *Infographic: Streaming Drives Global Music Industry resurgence*. Statista Infographics. Retrieved November 19, 2022, from https://www.statista.com/chart/4713/global-recorded-music-industry-revenues/

Global music report. IFPI. Retrieved from https://www.ifpi.org/wp-content/uploads/2022/04/IFPI_Global_Music_Report_2022-State_of_the_Industry.pdf

Note- first link in the references is the link to the dataset.