MUSIC DATA VISUALIZATION

Team Name: *'Symphony'*
Team Members: *Nidhi Vraj Sadhuvala*

ABSTRACT:

Music seems to be innate in humans, and we all hold it close to our hearts. It's a powerful weapon for uplifting people's spirits and brightening their days. Music is something that all cultures commonly enjoy. We all know speaking is an essential tool for communicating ideas or concepts, but music has an even stronger impact when compared to speaking, as it fulfills this function exclusively by fusing it with a range of instruments and vocal sounds that have a stronger impact to be implanted in one's brain. Sometimes when listening to such good music, dopamine is released when the song hits an emotional high point and the listener experiences chills or the shiver-inducing emotion of awe and exhilaration. Which gets us to the question, "What kind of music is most liked among listeners?", which serves as the basis for the study's title, "Why do we love music?".

A new upcoming musician may find it difficult to compose the ideal song that would be enjoyed by the widest possible audience given that there are more than 8 Billion people in the world, each with their own musical preferences. We want to provide insights to support these musicians in an effort to bring more of such music into this world. Since music is a fusion of many different sounds, our primary objective is to analyze the various scientific and technological musical elements that are responsible for producing good music that wins the hearts of most people. In order to better understand the types of music that people enjoy and the appropriate number of values to choose for a musical hit by an artist through various interesting visualizations, as well as what should be taken into consideration when composing music that is so adored by all, we would like to conduct our research into these crucial musical elements such as instrumentalness, valence, tempo, acousticness, loudness, speechiness, explicit content, etc. For examining the relationships between these various musical elements chosen by top artists, we have taken Top Hits on Spotify dataset from kaggle.

Why Spotify Dataset?

With the advancement of technology, the period of listening to music has changed from CDs to iPods, and then from iPods to phones. We are all familiar with the popular music program Spotify, which has amassed all of the records from the late 1990s to the mid 2000s in one location and has a sizable user base. We found Kaggle to be the greatest site in our hunt for where to go for the data that will work best for our research. We believe that the Spotify statistics from Kaggle would best represent our analysis.

INTRODUCTION:

BACKGROUND:

Many times, music has transported us to past memories by reminding us of certain people, boosting our spirits, and energizing us from a cheerful attitude to levels where we fully lose ourselves in a world where everything is calm and joyful without any thoughts. It has a great influence on our emotions and actions in a way that it can affect someone in both positive and negative ways. There are certain elements in music such as rhythm, harmony, melody and vocal sounds, but such sounds also contain a great deal of technical internal elements that might cause a listener's heart rate to rise. When analyzing a song, internal musical characteristics such as rhythm, tempo, beat strength, loudness, energy, speechiness, acousticness, instrumentalness, liveness, valence, and other properties of sound are taken into account.

It is nearly impossible to pinpoint what makes a song popular, regardless of the song's quality or the talent required to write and perform it. We've always been interested in what makes a song appealing to a broad audience as music fans. In order to determine the solution, we performed analysis to identify patterns and trends in the list of Spotify's Top Tracks from 2000 to 2019. We aim to study and visualize these components as part of our project to uncover some intriguing things about what makes good music. All of the musical element values considered by top artists based on the popularity they gained through some of the hits in their careers from the years 2000-19, we want to illustrate the insights through various visualizations that are helpful for new aspiring musicians in a way that can be used to accelerate their content for music. By examining the relationships between these various musical elements chosen by these top artists, depicting the reasons for popularity upon the chosen factors, studying the song functionality and choice in making the music, and contrasting it with the less popular songs to find the difference, we have advanced our research into the factors that contribute to the creation of a musical album.

MOTIVATION:

From our personal experiences, as foreign students studying far from home, we can at times feel lonely, under pressure, and stressed out about numerous things including our careers, jobs, assignments, and finals. The only way we can combat these feelings is to listen to some good music. By listening so, we feel relaxed and are able to focus on the task at hand rather than wander off to other unimportant topics. Many times, when we go out with friends to a venue that has some fantastic music and live performances, it often energizes the atmosphere and we can connect easily with others. In some cases, it may even be the only thing that two people have in common that allows them to engage in discussion. It has the potential to brighten the space and create memories that will last a lifetime.

Since music has had a significant impact on our lives, we are exposed to it every day. Music is usually heard during the commercial breaks when watching the morning news. We might listen to our favorite radio station while making supper, or while driving to work or school. Music continues to be at the heart of human civilization as a way to convey inner feeling, a way to celebrate life, and a symbol of enduring memories. The idea that music affects a person's life in various ways spurred the urge to study it and comprehend the various elements and variables at play.

Music has such a powerful effect that it transports us to another realm where it can make a positive difference in people's lives. Thus, producing high-quality music is essential, but in order to understand more about it, we must first completely understand music and search for its empowering influences. To accomplish this, we must look at the music composed by the most well-known artists and investigate the numerous factors that contributed to the song's popularity. This analysis of music-related data is important because, as already mentioned, music has a big impact on people. Additionally, it is essential because a lot of excellent music albums have the capacity to inspire and make people happy. We want to get such valuable insights and interesting facts through these visualizations in order to aid an artist in producing high-quality music.

OBJECTIVES:

Through this project, we aim to examine relationships and important elements that should be considered when composing music, such as song popularity and a number of other factors responsible, and observing the correlations of different elements. For everyone who aspires to compose good music, these visuals will be of great assistance. Here are a few of the inquiries we intend to investigate for the project and draw conclusions from:

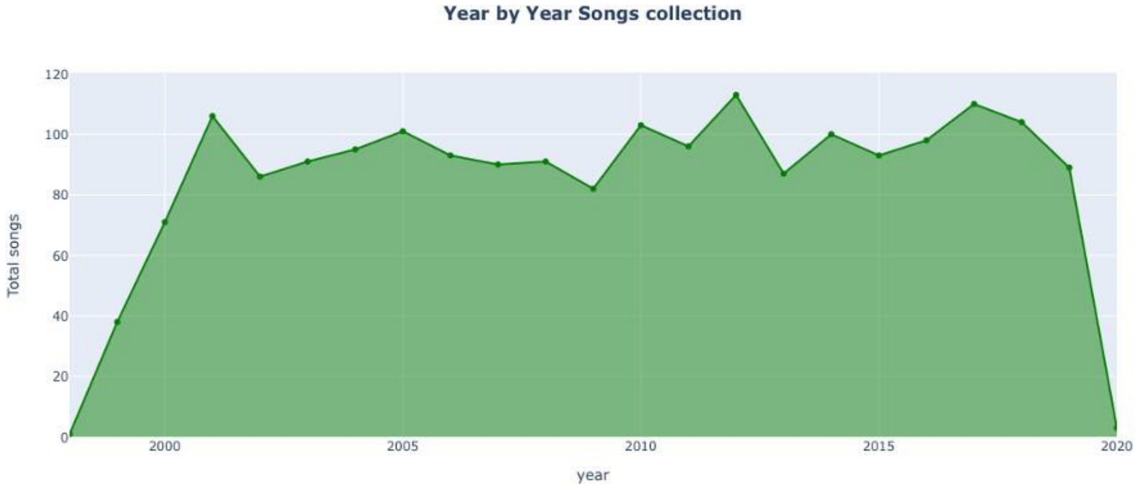
1. What musical elemental values to consider when composing good music and comparing the difference in values of top and least popular artists by taking into consideration the average musical elemental values.
2. Understanding the distribution of each variable over the whole dataset.
3. What is the count of the total number of songs produced per year and why did the number of musical hits decrease per year.
4. Understanding the relation of different genres with the song popularity and different other musical elemental values.
5. Deriving the relationship (correlation) between the different variables involved in the music and visualizing their relation over the years and producing better understandings in comparison with the existing methods.
6. Showcasing a hierarchical overview of different singers w.r.t genres they had produced songs in, along with different musical elemental values, based on popularity.
7. Investigating other features such as explicit content, melody content, duration of popular songs, and the percentage of top artists to find interesting insights.

EXISTING WORK/VISUALIZATIONS:

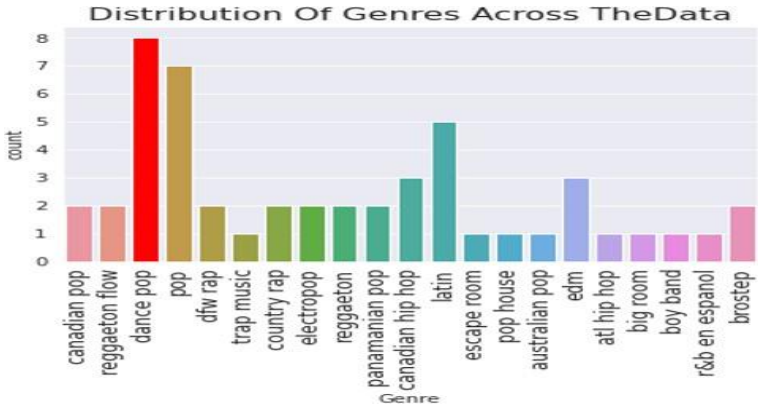
When investigating the data analysis and visualization techniques currently being used for music across various platforms, we came across a lot of intriguing visualizations that helped us understand the key insights and visualization techniques used. But after closely analyzing each of those representations, we came to the conclusion that they were more derived in the way they showed the tabular data that had already been established.

The following are some illustrations of these visualizations:

1) For instance, the picture below shows how simply two variables relate to one another. We may improve this representation by gradually adding many more musical elemental values for an easy comparison in the same graph.

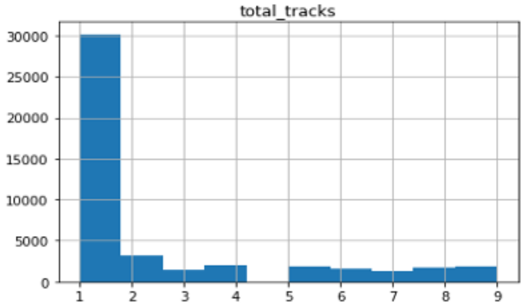


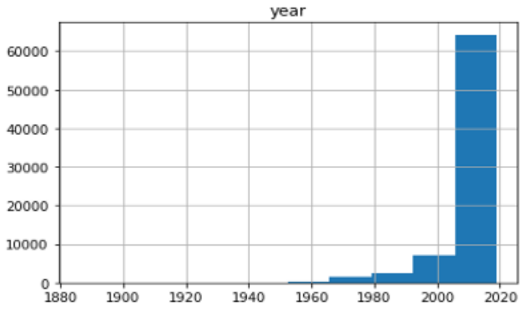
2) The following visualization demonstrates the count of the different musical genres across the dataset under study. It might be improved even more by either plotting a distribution graph or by using another plot that includes more genre-related elements that would explain how well-liked the songs are by an artist in comparison to music of different genres by other artists.



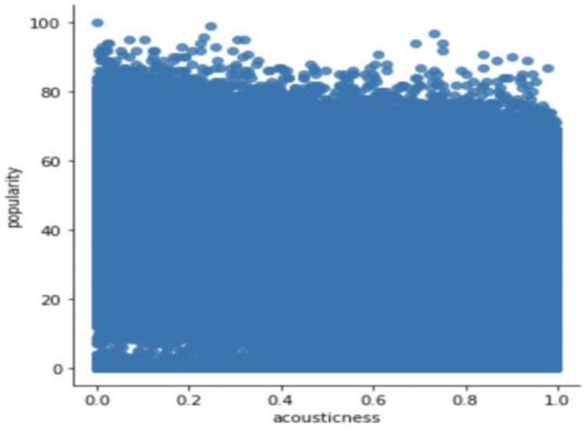
3) The graph below shows the overall number of tracks over time, but if you look closely, you'll notice that the second graph—which shows the number of songs each year—has a large bin size on its x axis and is difficult to grasp the exact year and the exact points it represents on y-axis. By

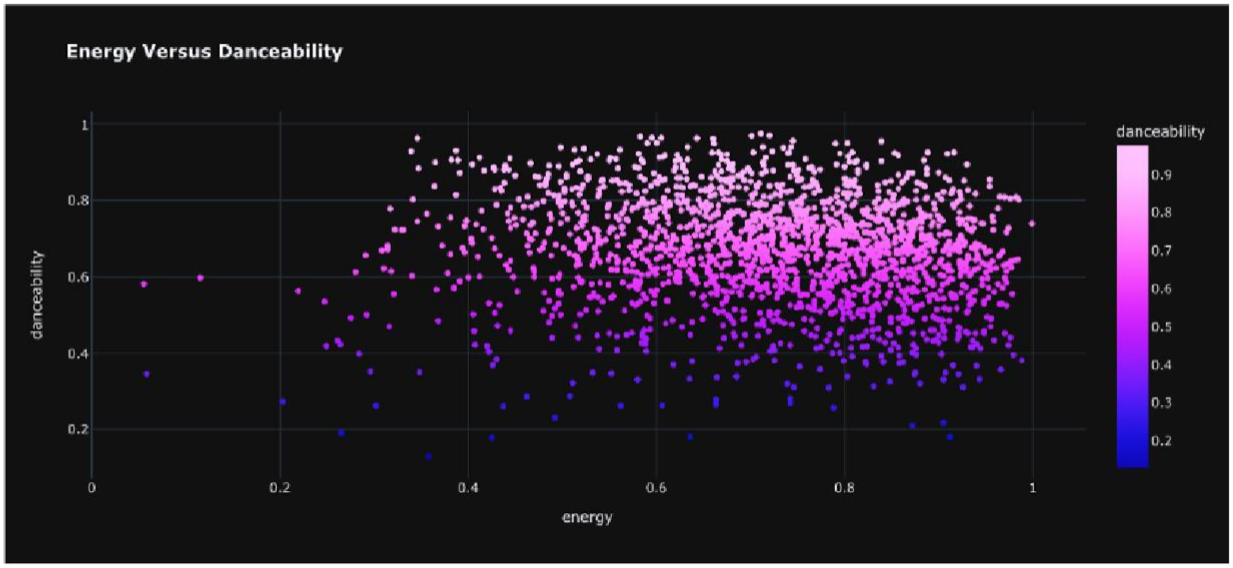
substituting a bar chart for the histogram, which will give the precise count for each year, we hope to make this graph easier to grasp.



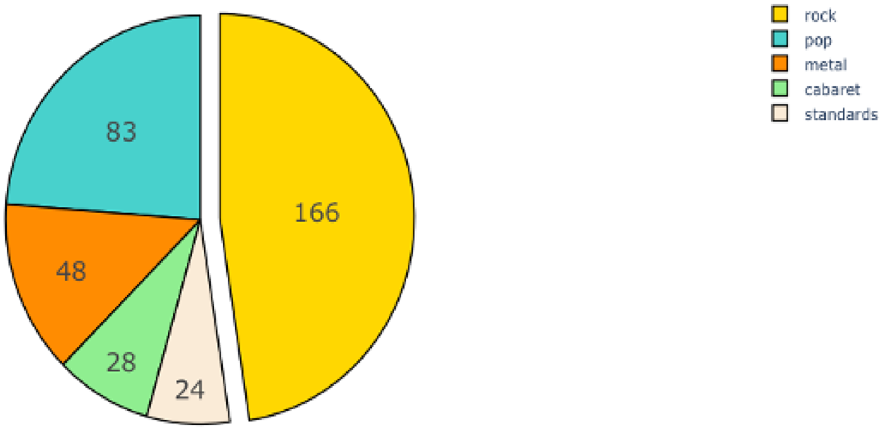


4) The correlation between two musical elemental factors is shown in the scatter plot visualization below. However, since the spread of the variables' values are closely clustered in this case, we believe that the scatter plot when plotting visualization between these two variables creates a problem with discretization of the values for this particular relation; as a result, it could be further improved by showing the analysis with less scatter points by plotting them over the relevant years ny popularity, which results in a clear understanding of the points and analysis of the visualization.





5) The pie chart in the following graph shows how popular certain musical genres are, yet there are more than five musical genres that we could analyze. By including the majority of the genres that are now available and showing their relationship with several variables and popularity associated, we can improvise this graph.



We can infer from the aforementioned visualizations that they mostly present tabulated data in the shape of graphs and offer relatively few insights and meaningful data. Therefore, we sought to enhance and show with the appropriate and effective interactive visualizations that would aid us in better comprehending the data and relationship between these variables, ultimately leading to the generation of better insightful conclusions. Also, according to our research, the majority of visualizations show the relationship between two different variables, but given the advancements in technology, we would like to show some visualizations that show the relationship between more than two variables making it multivariate analysis visualization. This would help us better understand the relationship between the variables and show it all at once.

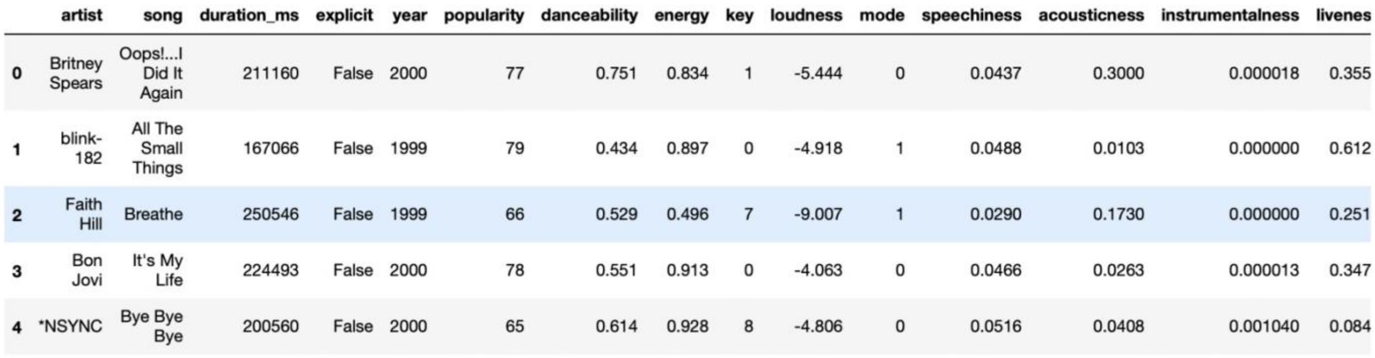
DATASET EXPLAINED:

In our search for the best dataset that is suitable for our analysis, we came across one of Spotify's data sets. As we mentioned above, Spotify is a widely used and more dependable platform and what other app could be better for choosing data other than the largest music streaming service used by almost everyone. We chose this dataset from Kaggle.

Below is a reference to the dataset's URL:
https://www.kaggle.com/code/varunsaikanuri/spotify-data-visualization/data

We chose this dataset because, as was already indicated, the elemental variables are the main defining characteristics we are seeking for in a dataset. This particular data set contains all the relationships and music elemental features that we require, along with details about the artists, genre, and the year the song was composed. We wish to examine these technical elemental properties of a piece of music. Selecting the suitable dataset will provide us with the justification and proof we require to back up our conclusions, allowing us to confidently defend them in the future and hence we believe that this is the appropriate and a tidy one. Without the right dataset, the chances of making mistakes and drawing incorrect conclusions increase significantly. We tried with a number of alternative datasets before settling on this one, and each one presented its own set of challenges. The data needed to be complete, with no components missing as this would cause the data to be disrupted. Lacking a precise set of data would make our visualizations less enlightening because we have encountered multiple problems, including difficult-to-read data values and challenges processing various data types and inconsistencies in the values.

The following image provides information about the data set, which comprises the columns:



	artist	song	duration_ms	explicit	year	popularity	danceability	energy	key	loudness	mode	speechiness	acousticness	instrumentalness	livenes
0	Britney Spears	Oops!...I Did It Again	211160	False	2000	77	0.751	0.834	1	-5.444	0	0.0437	0.3000	0.000018	0.355
1	blink-182	All The Small Things	167066	False	1999	79	0.434	0.897	0	-4.918	1	0.0488	0.0103	0.000000	0.612
2	Faith Hill	Breathe	250546	False	1999	66	0.529	0.496	7	-9.007	1	0.0290	0.1730	0.000000	0.251
3	Bon Jovi	It's My Life	224493	False	2000	78	0.551	0.913	0	-4.063	0	0.0466	0.0263	0.000013	0.347
4	*NSYNC	Bye Bye Bye	200560	False	2000	65	0.614	0.928	8	-4.806	0	0.0516	0.0408	0.001040	0.084

The data set includes information about the well-known songs and artists from the years 2000 to 2019 as well as the corresponding elemental values for each of the listed albums. It has a total of 2000 rows and 18 columns.

DATA DESCRIPTION:

The dataset has 4 of the columns containing categorical values, while the rest are numerical. The data for these columns is stored in a tabular manner, which is one of the most crucial factors in ensuring that this is a tidy dataset, meaning that every column is a variable and all the observations are stored in rows in a structured manner.

The description of the variables in the dataset is as follows:

Categorical Variables:

- artist: Name of the artist
- song: Name of the song
- explicit: Represented as a boolean value; one or more of the following elements present in the lyrics or content of a song or music video and may be seen as offensive or inappropriate for young listeners
- genre: Category that a music belongs to

Numerical Variables:

- duration_ms: Duration of the song in milliseconds
- Year: Release year of the song
- popularity: The higher the value, the more popular the song is
- danceability: Based on a variety of musical qualities, such as tempo, rhythm stability, beat strength, and general regularity, danceability explains how appropriate a track is for dancing. The least danceable value is 0.0, and the most danceable value is 1.0.
- energy: Energy represents a perceptual measure of intensity and activity and ranges in value from 0.0 to 1.0.
- key: Using the traditional Pitch Class notation, integers correspond to pitches. Value -1 is returned if no key was found.
- loudness: The overall decibel level of a track (dB). In order to compare the relative loudness of tracks, loudness ratings are averaged over the entire track. Typical values are between -60 and 0 db.
- mode: the type of scale from which its melodic content is formed, is indicated by the term "mode." Major is symbolized by 1 and minor by 0.
- speechiness: Speechiness is the ability to identify spoken words in music. The attribute value is closest to 1.0 for recordings that are mostly speech-like or involving many spoken words having values above 0.66. Tracks that may include both music and speech, either in portions or layered, are described by values between 0.33 and 0.66, encompassing situations like rap music. Values less than 0.33 most likely refer to tracks that are not speech-like, like music.
- acousticness: A scale of 0.0 to 1.0 used to determine if a track is acoustic. 1.0 denotes a high degree of assurance that the track is acoustic.
- instrumentalness: Determines whether a track is instrumental or vocal-free. Sounds like "ooh" and "aah" are regarded as instrumental in this situation. Tracks that are spoken word or rap are vocal. The likelihood that a track is vocal-free increases as the instrumentalness value approaches 1.0.
- liveness: Greater liveness numbers indicate a higher likelihood that the song was performed live. A score greater than 0.8 indicates a high probability that the music is live.

- valence: Valence is a scale from 0.0 to 1.0 used to describe how positively melodic a track sounds. Those with a high valence sound happier, cheerier, and more euphoric, whilst tracks with a low valence are more depressing (eg: sad, depressed, angry).
- tempo: The estimated overall tempo of a track, expressed in beats per minute (BPM). Tempo, which in musical terms refers to a piece's pace or tempo, is directly related to the length of an average beat.

DATA PROCESSING AND CLEANING:

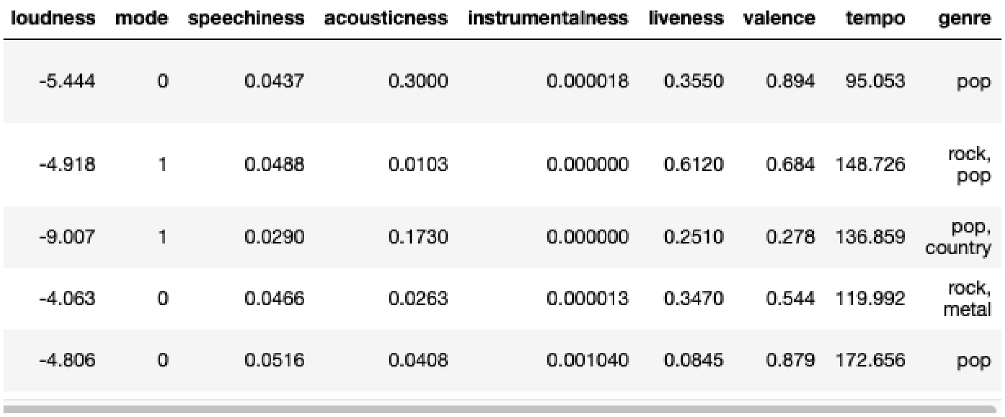
The data involved multiple data types for certain columns like:

Strings: Artist, Song and Genre
Boolean: Explicit
Decimal values: Danceability, Energy, Speechiness, Acousticness, liveness, Valence, Tempo
Negative values: Loudness
Scientific Notation values: Instrumentalness

Step 1: Removing the null values and scaling the columns to transform the data into a common scale

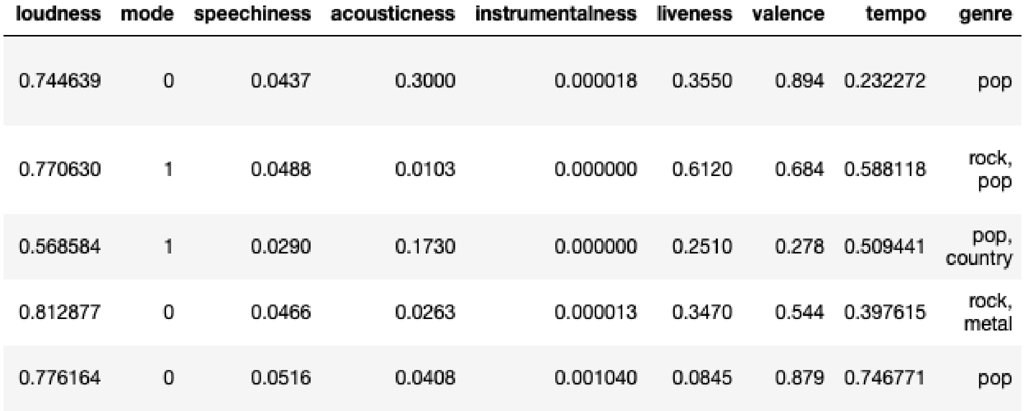
After initially skimming through the dataset, we looked for null values and found that there weren't any, which made the data set simpler and more easy to work with. However, there are few columns in our dataset that have negative and scientific notation values, and often, it becomes difficult to assess trends and patterns and compare features or columns when a dataset contains values for several columns at vastly different scales. Therefore, it is sometimes vital to ensure that the numerical columns of a dataset are changed to a common scale in cases when all the columns have a noticeable variance in their scales. We used the MinMaxScaler from sklearn to scale “tempo” and “loudness” columns to a value between 0 and 1.

Before Scaling: loudness and tempo values before scaling



loudness	mode	speechiness	acousticness	instrumentalness	liveness	valence	tempo	genre
-5.444	0	0.0437	0.3000	0.000018	0.3550	0.894	95.053	pop
-4.918	1	0.0488	0.0103	0.000000	0.6120	0.684	148.726	rock, pop
-9.007	1	0.0290	0.1730	0.000000	0.2510	0.278	136.859	pop, country
-4.063	0	0.0466	0.0263	0.000013	0.3470	0.544	119.992	rock, metal
-4.806	0	0.0516	0.0408	0.001040	0.0845	0.879	172.656	pop

After Scaling: loudness and tempo values after scaling are now comparable to other variables



loudness	mode	speechiness	acousticness	instrumentalness	liveness	valence	tempo	genre
0.744639	0	0.0437	0.3000	0.000018	0.3550	0.894	0.232272	pop
0.770630	1	0.0488	0.0103	0.000000	0.6120	0.684	0.588118	rock, pop
0.568584	1	0.0290	0.1730	0.000000	0.2510	0.278	0.509441	pop, country
0.812877	0	0.0466	0.0263	0.000013	0.3470	0.544	0.397615	rock, metal
0.776164	0	0.0516	0.0408	0.001040	0.0845	0.879	0.746771	pop

Step 2: Removing duplicates

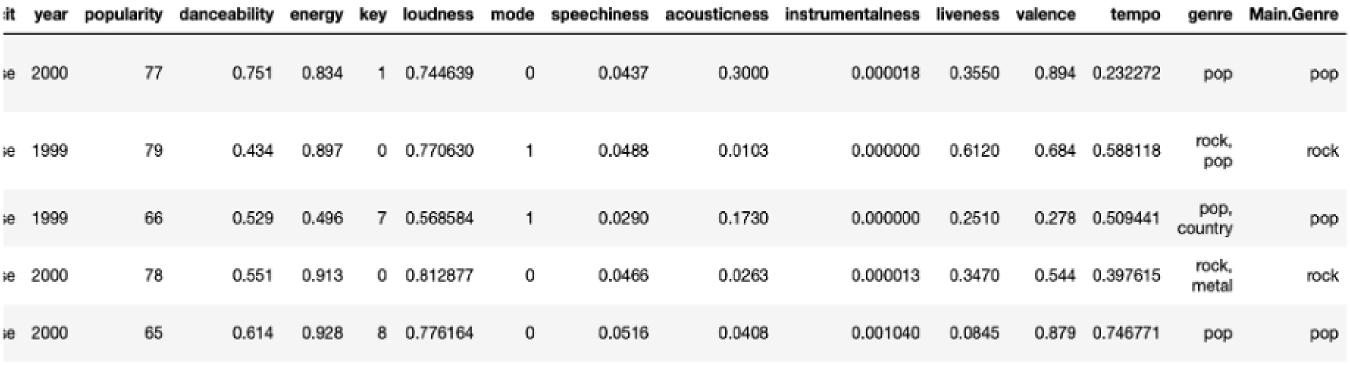
Additionally, we looked for duplicates in the dataset and found 59 of them. We then deleted these duplicate rows.

Step 3: Converting the duration of the song from milliseconds to minutes

The duration of a song is described in a particular column in the provided dataset. This column was originally written in milliseconds, but we changed it to minutes because minutes are the more common norm and people are more accustomed to seeing this. We converted this column from milliseconds to minutes in order to connect with the actual situation. This helps the audience perceive and understand the information better.

Step 4: Extracting and representing the main genre for an album

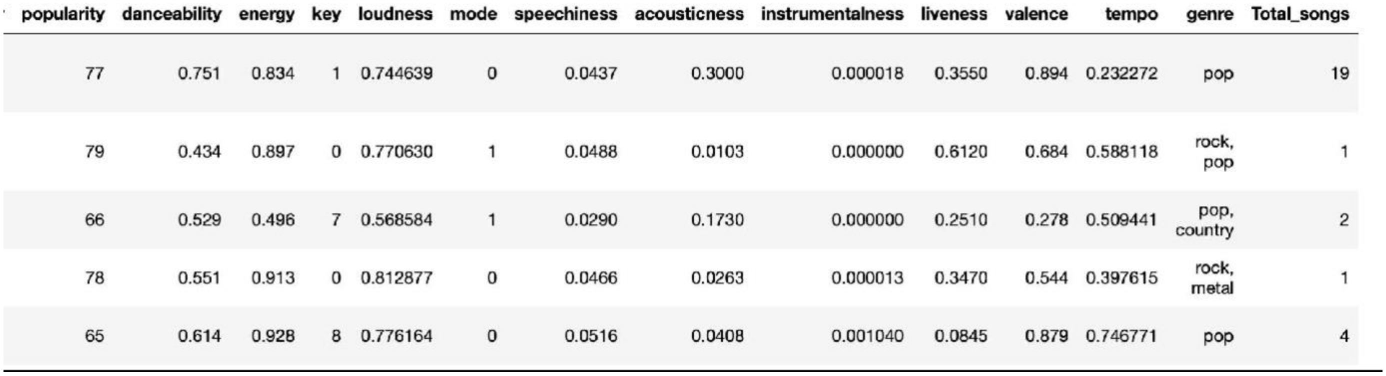
A certain genre can be applied to any musical album. Because the genres indicated for each album in the selected dataset involve more than one category and make it challenging to evaluate the genres, we have chosen to provide each album with its primary genre by adding an additional column to the original dataset called the Main Genre as displayed below.



it	year	popularity	danceability	energy	key	loudness	mode	speechiness	acousticness	instrumentalness	liveness	valence	tempo	genre	Main.Genre
se	2000	77	0.751	0.834	1	0.744639	0	0.0437	0.3000	0.000018	0.3550	0.894	0.232272	pop	pop
se	1999	79	0.434	0.897	0	0.770630	1	0.0488	0.0103	0.000000	0.6120	0.684	0.588118	rock, pop	rock
se	1999	66	0.529	0.496	7	0.568584	1	0.0290	0.1730	0.000000	0.2510	0.278	0.509441	pop, country	pop
se	2000	78	0.551	0.913	0	0.812877	0	0.0466	0.0263	0.000013	0.3470	0.544	0.397615	rock, metal	rock
se	2000	65	0.614	0.928	8	0.776164	0	0.0516	0.0408	0.001040	0.0845	0.879	0.746771	pop	pop

Step 5: Adding a new column Total Songs in order to average other variables and prevent data distortion

Since just a few musicians have generated numerous songs and these artists have frequently appeared in the dataset, we included a new column called "Total Songs" that is calculated by adding the number of songs produced by each artist.Accordingly, when we sought to use this popularity measure to gauge the popularity of all artists, all of the popularity values were added up for the number of times an artist's name appeared in the dataset. Each artist is assigned a popularity value for each song. This popularity value summarization could produce inaccurate conclusions. When we add up the popularity numbers for each artist, the least popular artist will end up being the most popular one, however this is the wrong approach to follow and could mislead the audience. For instance, an artist may have produced 5 songs with a popularity below 50 and another artist with 2 songs with a popularity above 80. In order to change our dataset, we determined the total number of songs created.As a result, we changed our dataset by estimating the total number of songs each artist has recorded. We then used this information for various interpretations using the average of the variables.



popularity	danceability	energy	key	loudness	mode	speechiness	acousticness	instrumentalness	liveness	valence	tempo	genre	Total_songs
77	0.751	0.834	1	0.744639	0	0.0437	0.3000	0.000018	0.3550	0.894	0.232272	pop	19
79	0.434	0.897	0	0.770630	1	0.0488	0.0103	0.000000	0.6120	0.684	0.588118	rock, pop	1
66	0.529	0.496	7	0.568584	1	0.0290	0.1730	0.000000	0.2510	0.278	0.509441	pop, country	2
78	0.551	0.913	0	0.812877	0	0.0466	0.0263	0.000013	0.3470	0.544	0.397615	rock, metal	1
65	0.614	0.928	8	0.776164	0	0.0516	0.0408	0.001040	0.0845	0.879	0.746771	pop	4

VISUALIZATION METHODS EXMPLOYED:

The different visualization techniques we have used to conduct our analysis are as listed below:

- Heat Map
- Histogram Distribution Plot
- Polar Line Plot
- Radar Plot
- Donut Chart
- Pie Chart
- Swarm Plot
- Jittered 1D Scatter Plot
- Scatter Plot
- Tree Map
- Box Plot
- Violin Plot
- Polar Scatter Plot
- Line Plot
- Funnel Plot
- Sunburst

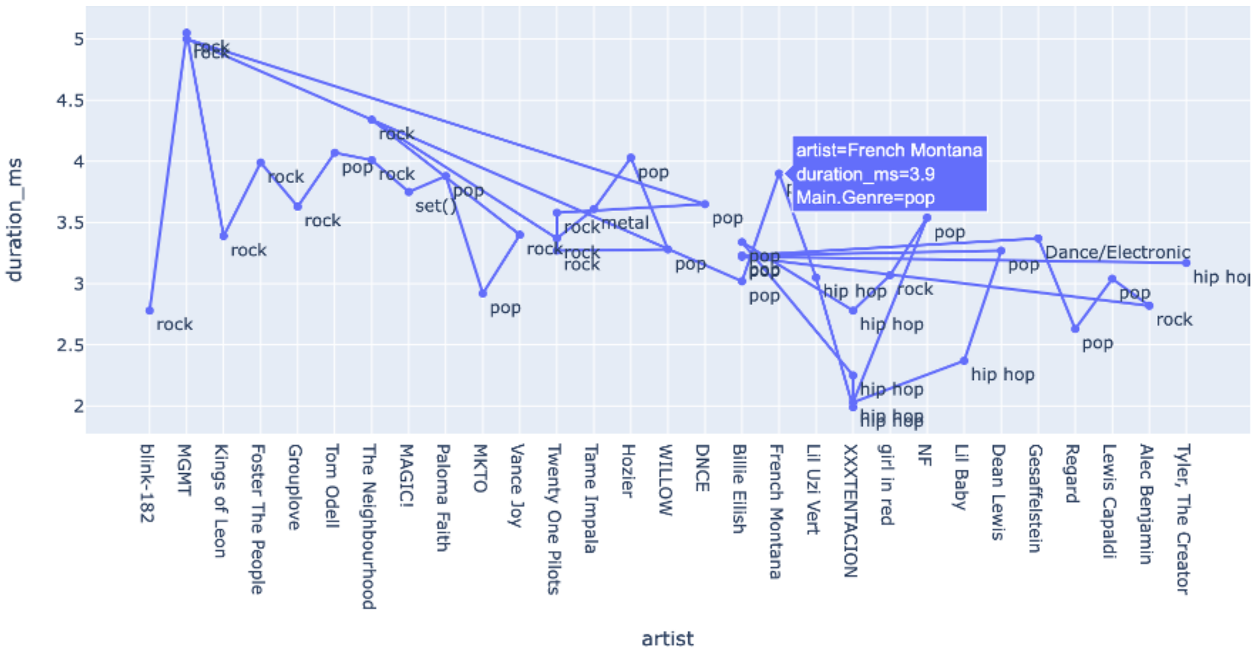
At first, we plotted various visualizations in connection to the different columns in the dataset to try to comprehend them. The dataset primarily consists of the fundamental values that go into making a musical album, which are represented in several numerical data types, as was previously indicated in the data description. We have generated a seaborn histogram plot along with a normal distribution for all the songs in the data to help us understand the musical elemental values and how the values of the variables are distributed. Then we went on to predict distinct connections between the elemental values where the complete dataset is taken into account. The relationship and average value utilized across by many performers for numerous songs have been displayed on a single plot using visualization techniques like the line polar and sunburst.

Furthermore, in order to comprehend the elemental values and use it to get insightful knowledge, we analyzed the popularity variable across the data using various visualizations, such as the scatter plot, pie chart, bar graph and so on, which we then used in the following visualizations to show various connections between musical components taken by an artist based on popularity. Since producing insightful analysis and practical knowledge about popular music is our primary goal.

In order to experiment about representing data in the best suited visualizations, we have plotted several graphs such as the tree map for hierarchical overview, the bar plot, the sunburst etc for drawing insights that collectively help in the overall analysis.

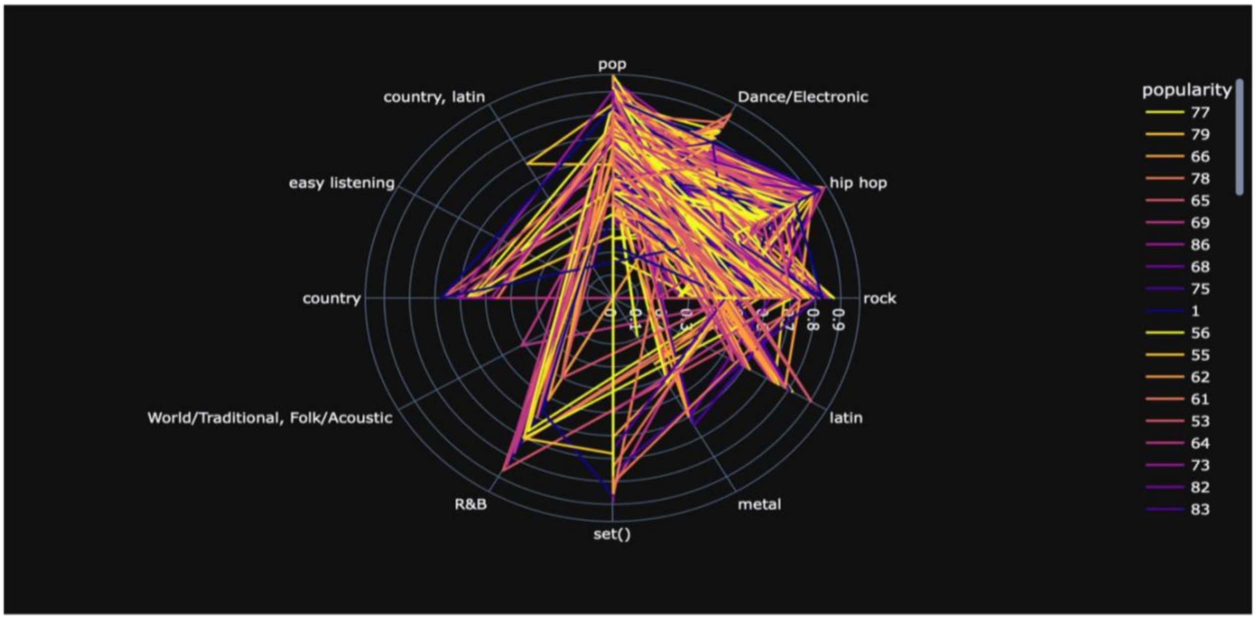
FAILED EXPERIMENTS:

1) When experimenting with the construction of visualizations across multiple relations, we discovered that not all representations are suited for drawing visually perceptive and meaningful insights. We finally selected the ideal visualization graphs for finding specific relationships between variables after making these mistakes and experimenting. For instance, in order to estimate the most well-known musicians' average song length in relation to the genre, we first produced projections using the line chart shown below:



The reading of the graph is complicated as the data points are scattered and it may not be the best chart for the audience in deriving important insights, making it extremely challenging to determine the relationship from the aforementioned plot. We have chosen the bar plot as a replacement, as shown below in the visualizations section for the provided relation, in order to ease the aforementioned complexity of finding data points. The bar plot depicts the same data without overlapping on one another in addition to including popularity as a third metric to aid in our analysis.

2) Additionally, in order to evaluate the danceability metric, we first built a graph similar to the one below to show the link between the variables of various genres based on popular demand:



The above representation is complex because the lines cross over into another genre, making it difficult to determine which lines are pointing to which genres. We decreased the visual complication in understanding the complexity for various genres by changing the lines on the polar to a scatter polar in order to make it more visually appealing.

RESULTS AND INSIGHTS:

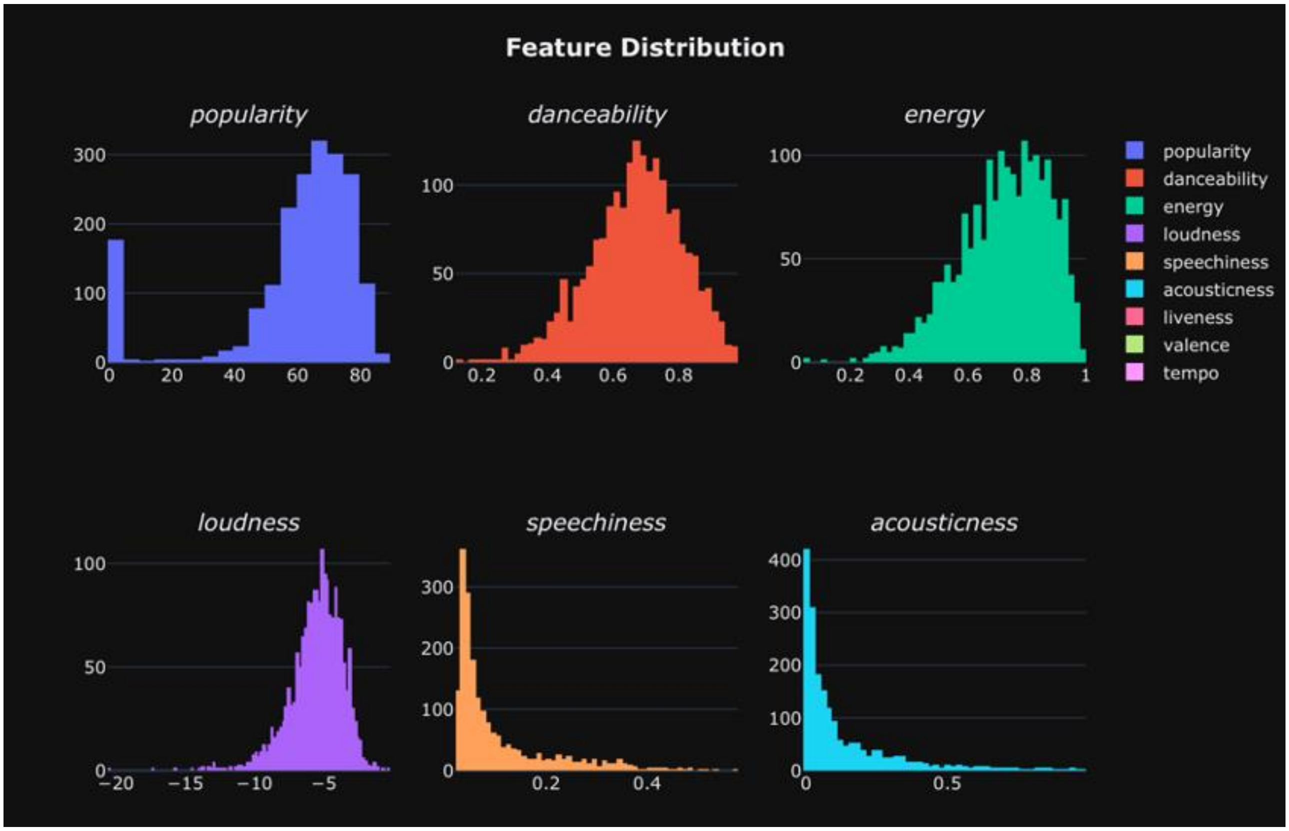
Visualization 1: Heat Map

The many color-coded representations of the data in the Heat Map allow us to more easily see the volume and existence of variables within the data. This visualization was created to direct the audience to the areas of the graphic that will help them ignore and find the relationships between variables. The stronger the correlation between the two variables, the larger the number and darker the hue. The closer the correlation is to 1, the more positively associated they are; that is, as one rises, the other does as well, and the stronger this relationship is. Hence, all the red and orange colored values are more strongly related such as energy and loudness have a value of 0.65. There is no linear trend between the two variables if the values are closer to zero or pointing to negative, but instead of both variables rising, one will fall as the other increases. Acousticness has the lowest value of the energy, which is -0.45, indicating that they are negatively correlated. As we can see, the accompanying heat map uses heat to represent each variable in the selected dataset. It was utilized to forecast the relationship between these two variables in the upcoming visualizations.

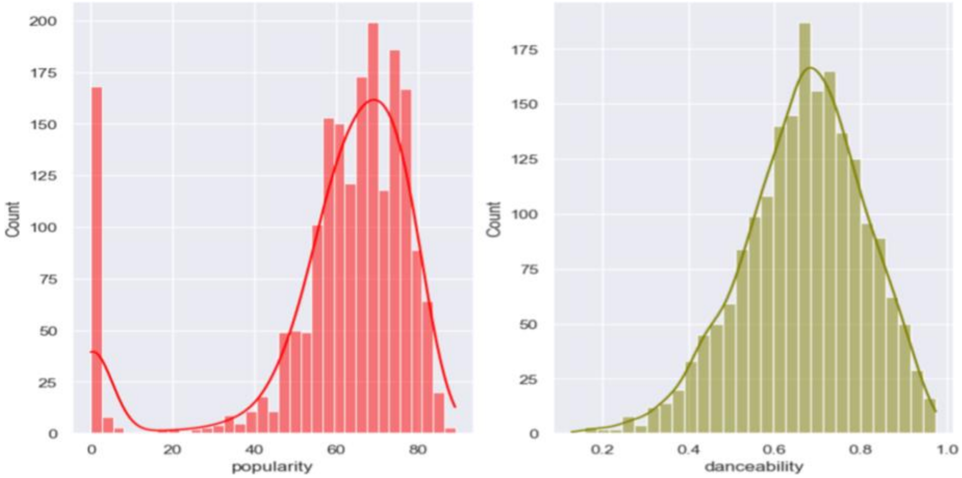
Visualization 2: Histogram Distribution of each feature

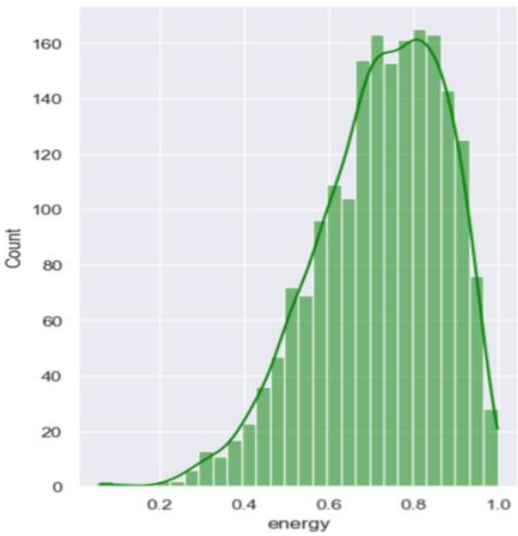
We modified the below existing graph and made a better suited visualization:

Existing:

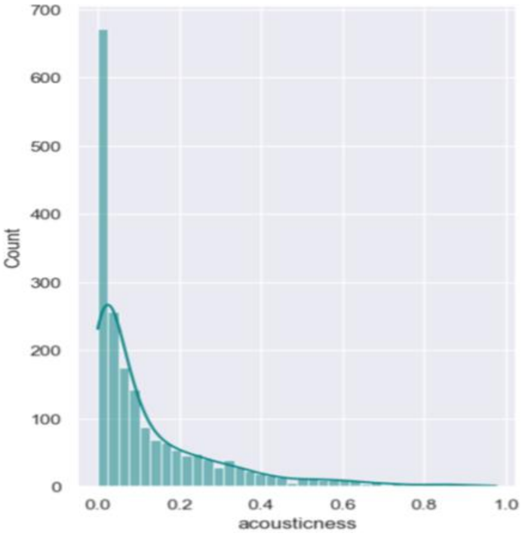


As seen in the graph above, the various elemental variables are depicted using a histogram plot throughout the full dataset. Although there are many different values represented in the image above, it is difficult to determine the median values using this graph. In order to further enhance the aforementioned visualization, we added a normal distribution to the plots, which could be used to forecast the median value for the fundamental variables, as seen below:

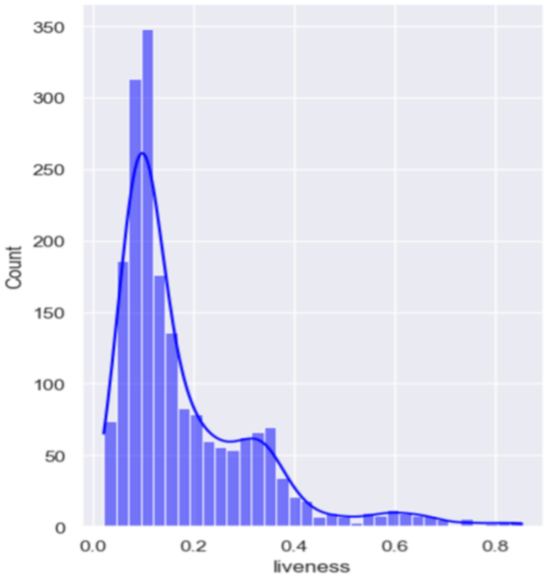




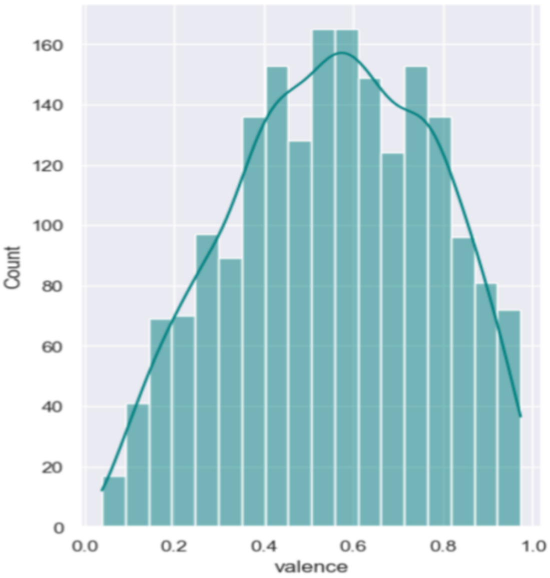
Count
0
20
40
60
80
100
120
140
160
0.2
0.4
0.6
0.8
1.0
energy



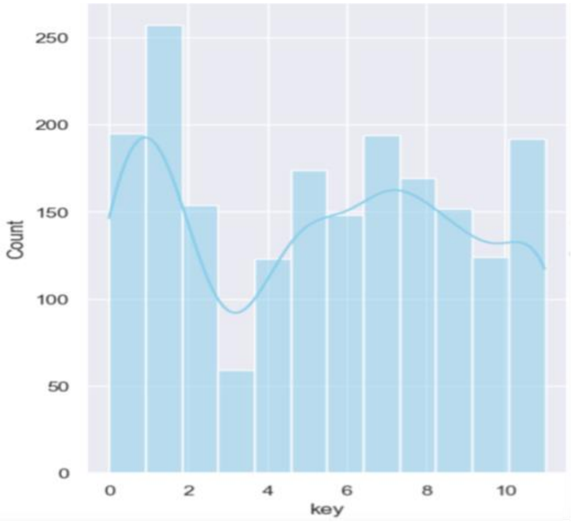
Count
0
100
200
300
400
500
600
700
0.0
0.2
0.4
0.6
0.8
1.0
acousticness



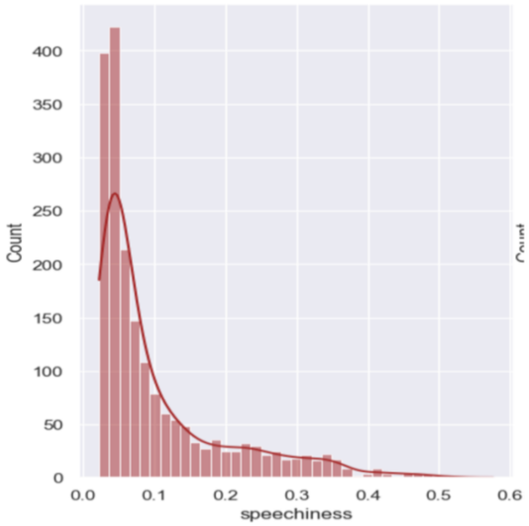
Count
0
50
100
150
200
250
300
350
0.0
0.2
0.4
0.6
0.8
liveness



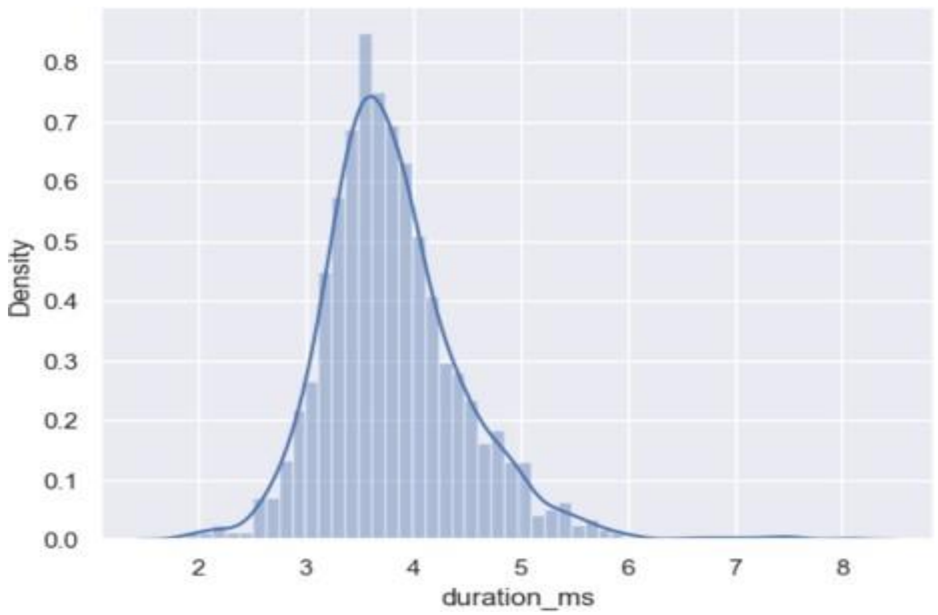
Count
0
20
40
60
80
100
120
140
160
0.0
0.2
0.4
0.6
0.8
1.0
valence



Count
0
50
100
150
200
250
0
2
4
6
8
10
key



Count
0
50
100
150
200
250
300
350
400
0.0
0.1
0.2
0.3
0.4
0.5
0.6
speechiness
Count

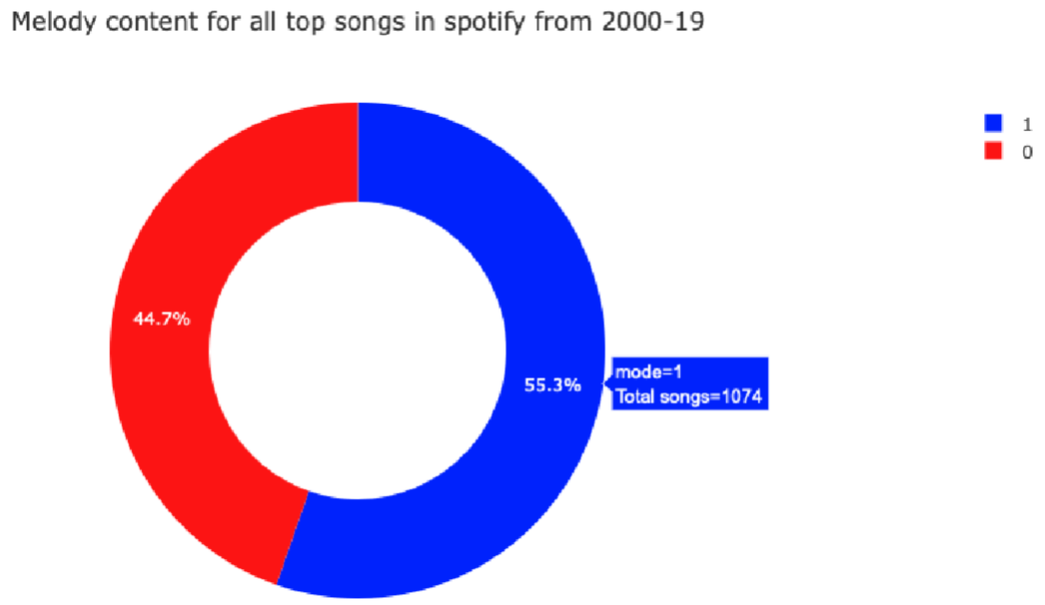


Using reference to "Module 09" from the data visualization course, we have implemented the idea of plotting with histogram and KDE combined in the following visualization. The histogram is uniformly distributed across the dataset for all the musical variables. This has made it easier for us to comprehend how the variables are distributed across the entire data set.

Insights drawn:

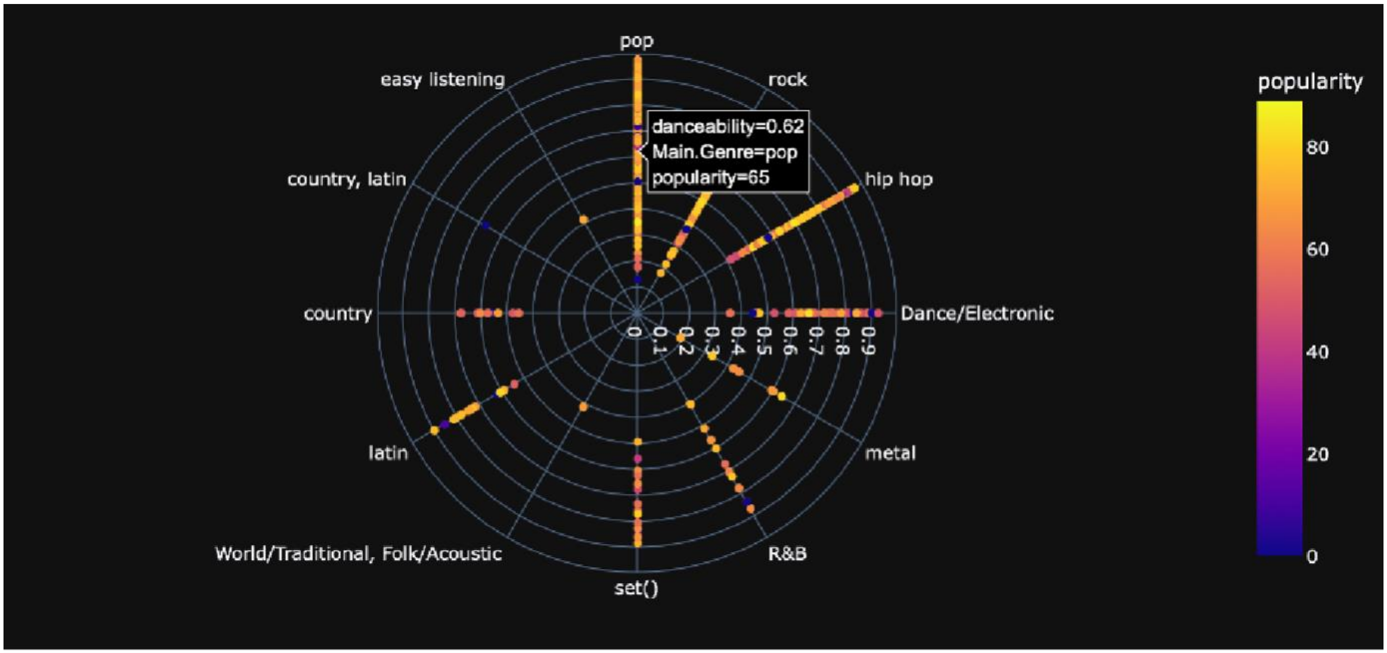
- When we look at the duration plot, for instance, we can see that this is a symmetric distribution and for such graphs mean, median, and mode are the same and this point is between 3 and 4, which indicates that the majority of the songs are between 3 and 4 minutes in length.
- Similar to duration plot, valence, and danceability, have almost normal distribution and we can conclude that mean, median and mode for these graphs are 0.6 and 0.7 respectively. These median values are useful when taking these components for composing music into consideration.
- The right-skewed distribution of speechiness, liveness, and acousticness explains why the mean is above the median and why the majority of values are centered on the left. When we look closely at the speechiness graph, we can see that artists chose fewer speech-like songs in the dataset because there are less values over 0.3, and the maximum speech-like songs should be about 0.66. When measuring liveness, we can infer that the majority of songs are recorded rather than live performances because the majority of values fall between 0.0 and 0.2, and a higher value would suggest that there is a greater likelihood of artists playing live. Very few songs are highly acoustic.
- Popularity and Energy variables have left-skewed distribution meaning that most of the artists have popularity values ranging between 50-80 and most of the energy values are between 0.7 to 0.9.
- The Key variable doesn’t follow a normal distribution and hence it’s hard to draw any insights from this.

Visualization 3: Donut Chart to depict the proportion of Melody content



We know that there are many distinct types of music, one of which is referred to as the melodic, and that there are numerous aspects that affect the popularity of music. Since everyone is aware of how calming and uplifting melodic music is, we wanted to test this to see if it is something that most people find appealing. We therefore generated a donut chart that shows the percentage of melodic content and non-melodic content over the list of songs in the dataset is a compilation of the top hits from the years 2000 to 2019. Since the proportion of melodic to non melodic songs is roughly equal, we can deduce from careful observation that while melodic songs don't necessarily contribute significantly to a song's popularity overall, but in general when examined closely melodic songs stand more popular.

Visualization 4: Polar Scatter Plot



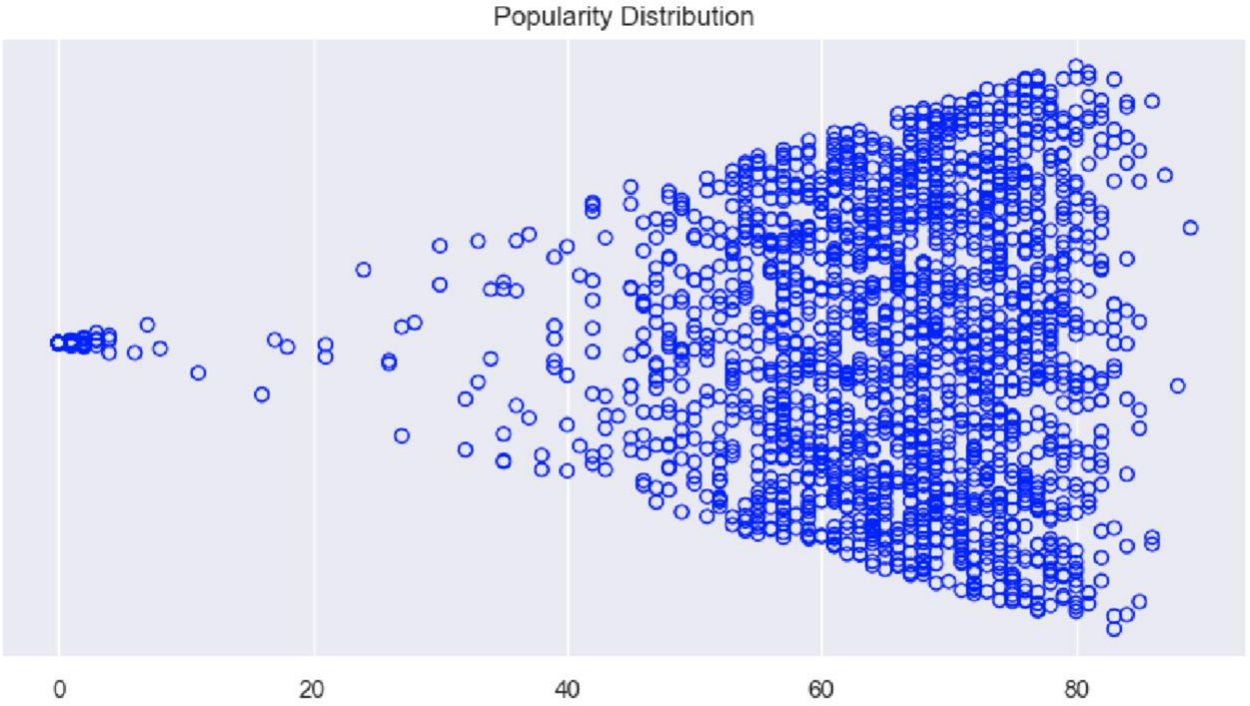
Every song is classified into different genres and each genre has a different impact upon an individual. Everybody has their own favorite music genre; consequently, it is crucial to comprehend the popularity of the music relative to the genre when an artist is first composing the music. When attempting to develop a visualization highlighting the genre's popularity, we also wanted to incorporate the danceability into the image and produce a multivariate analysis visualization, which led to the creation of the polar scatter plot you see above.

As observed from the above visualization we can see that the population is distributed over different genres considering the danceability as the radius. Additionally, we can see that the pop genre has the greatest number of data points, concluding that it is the most well-liked genre among individuals with a high danceability value, followed by rock, hip hop, and dance electronic as seen above. We can also see that other genres, including traditional, folk, and latin, are not particularly popular with people. This implies that compared to other music genres, this one may not be as well known to the general audience.

Visualization 5: Jittered 1D Scatter Plot

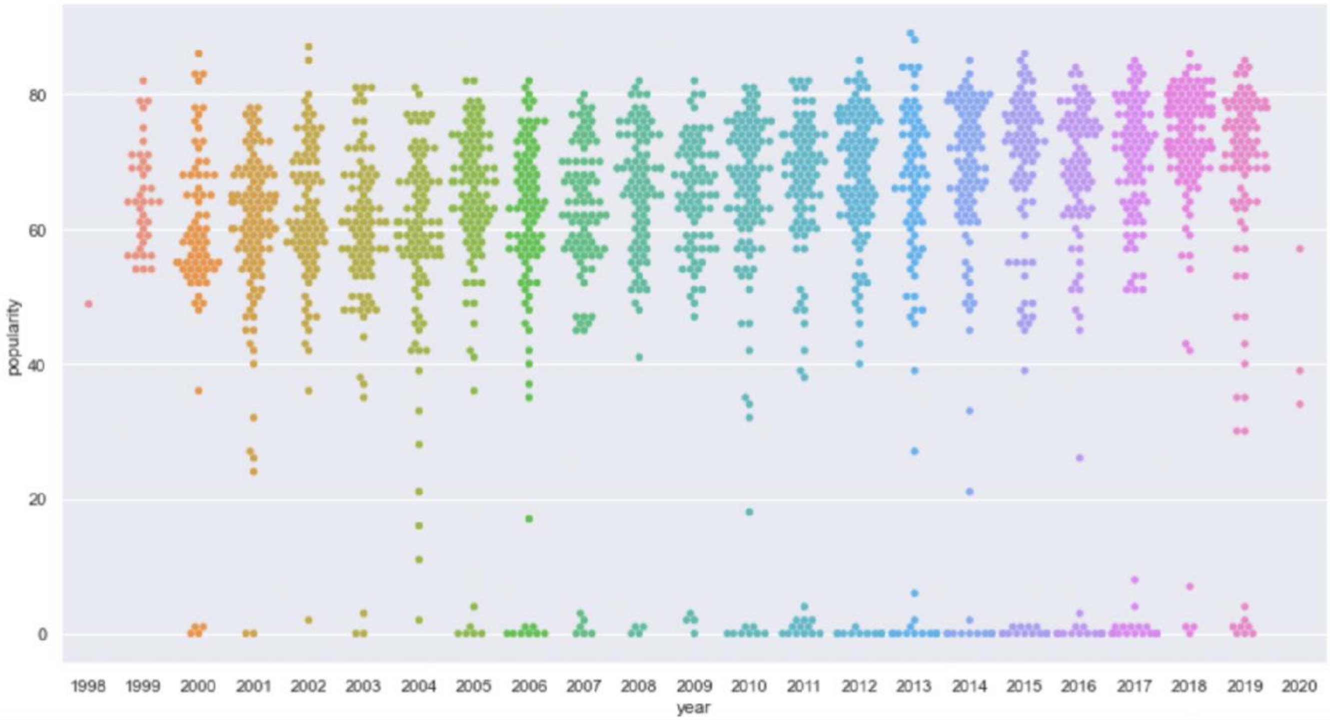
In order to find the insights and elemental values responsible for popular music, we have analyzed the popularity column through which we have gained certain insights:

- We have observed that the most popularity ranges from the values between 60 and 80 therefore the average value of the popularity is showcased here.
- This illustration makes it clear that the most well-known artists in the dataset have popularity values greater than 80. As a result, taking into account the artists whose popularity is near to or above 80 would place their popularity at the top, and taking into account their elemental values would aid in obtaining necessary important information.

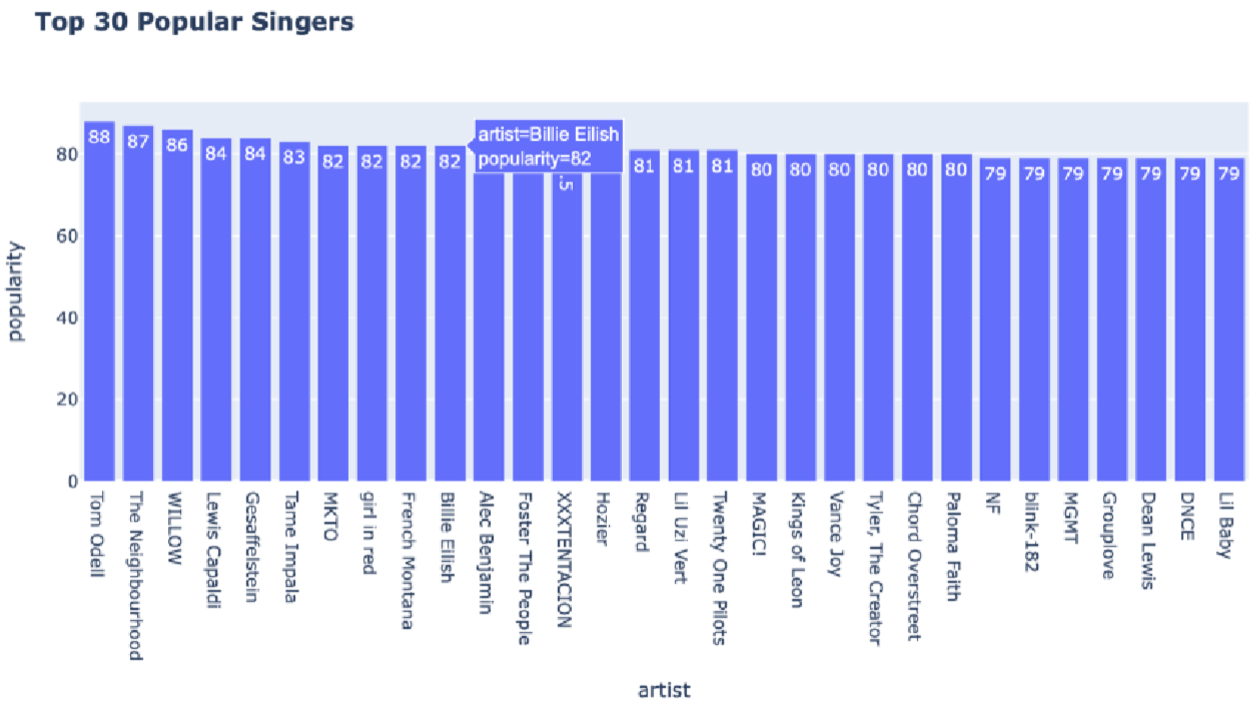


Popularity Distribution over the years:

We can observe that most of the popularity distribution is between 60 to 80 for all the years and the popularity density of the data points had increased over the years.

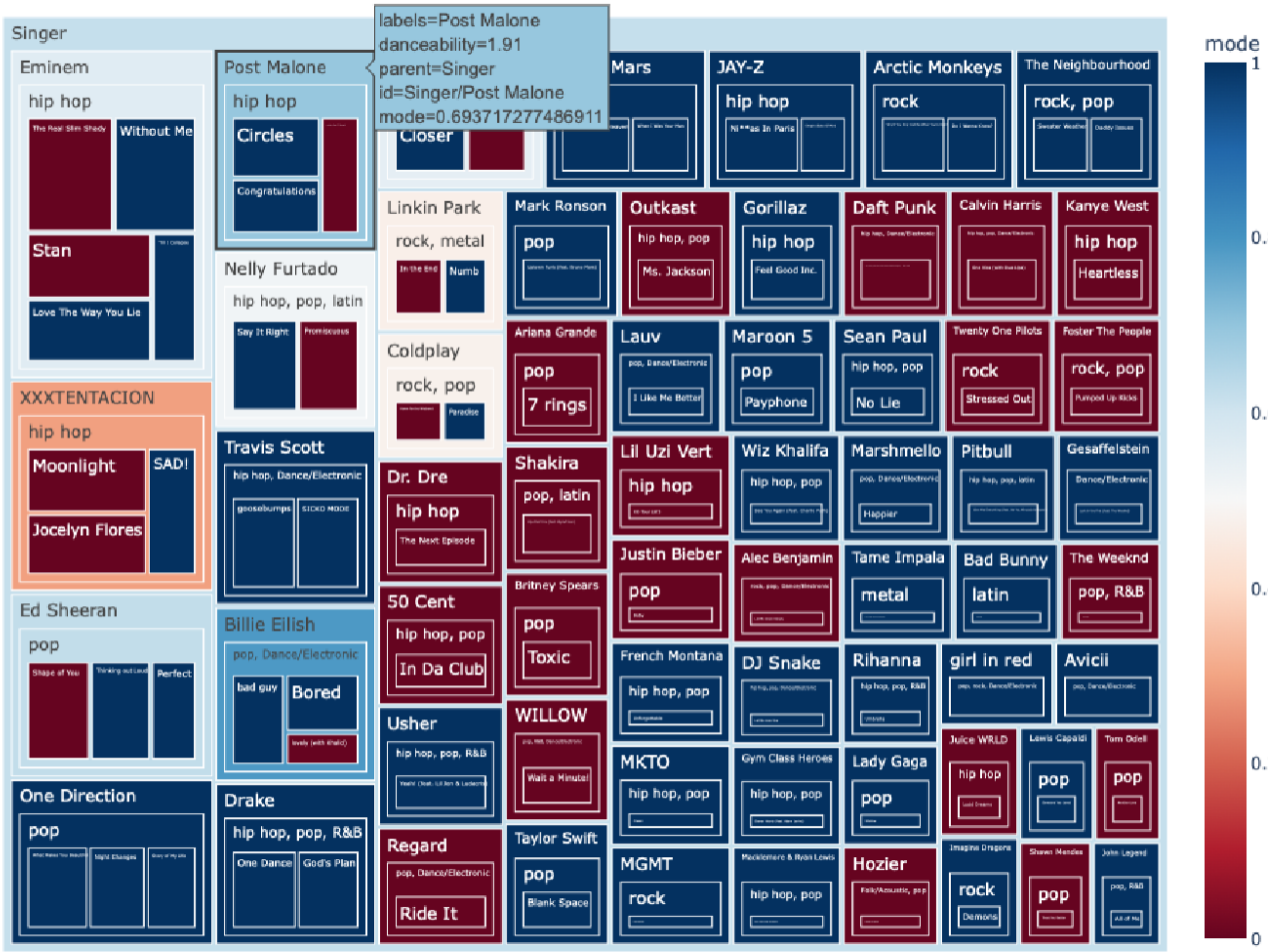


Visualization 6: Bar Chart depicting Top 30 Popular Singers



Following the popularity distribution plot, we discovered that the top artists fall within a range near or above 80. With this knowledge in mind, we set out to identify the top 30 most well-known musicians and examine their fundamental musical characteristics by taking the average of the popularity values of each artist. By doing this, we hoped to gain knowledge that would aid an upcoming musician in creating popular music by taking these components into account, which is the main goal of this project. The top 30 Spotify artists from 2000 to 2019 are displayed in the above bar plot visualization, which was created using the plotly library. The Top most popular artist is “Tom Odell” followed by “The Neighbourhood”, “WILLOW”, etc. respectively in the order mentioned.

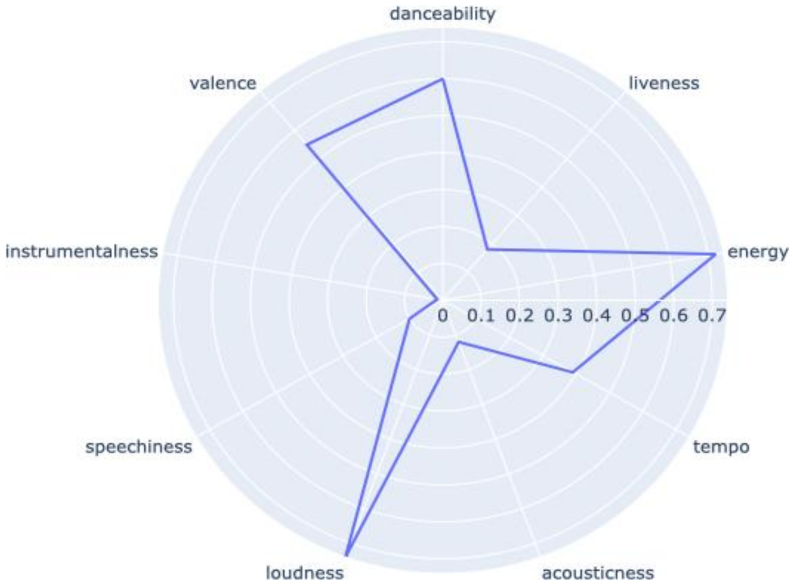
Visualization 7: Tree Map depicting hierarchical overview of each singer



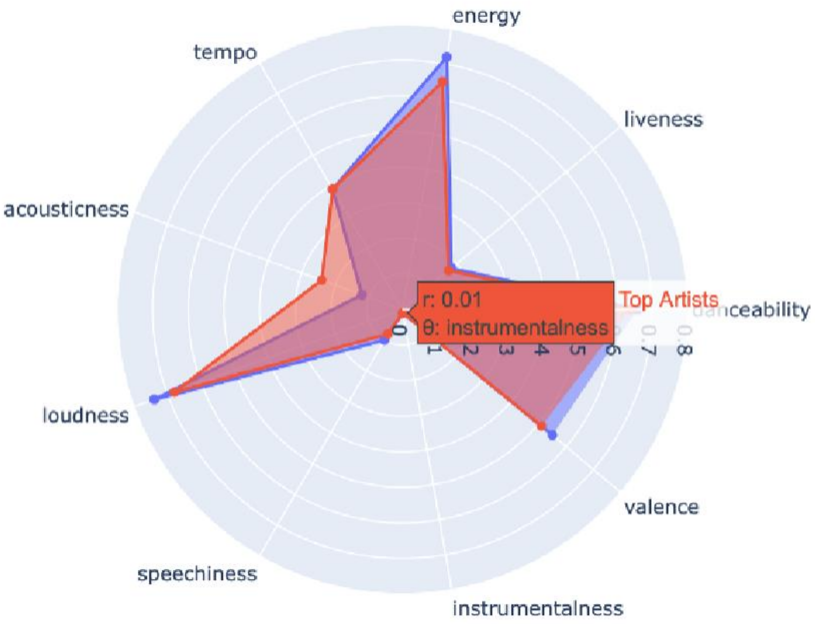
In order to make the data set easier to grasp, we sought to design a visualization that accurately depicts as much information as possible for all top artists based on popularity value > 80. We have determined that the Tree map would be the best fit for this after conducting numerous experiments and investigations to see which visualization would best display this information. The plotly library was chosen because it would accurately depict the precise numbers and data when you hover over the display. The top 30 musicians are shown in this image along with the genre they had composed music in, along with the danceability and mode values.

Visualization 8: Radar chart to compare key data points

Radar chart depicting the range of values of all features for the overall dataset



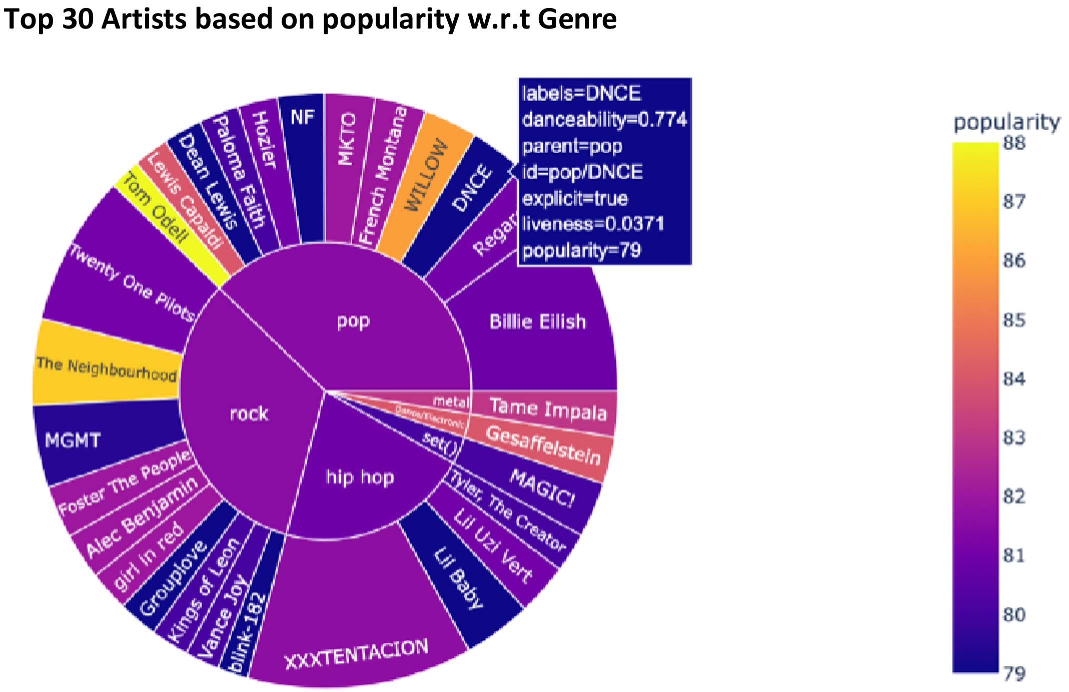
Radar chart depicting the top artists elemental values range infused on the overall dataset range



We constructed a list of the top 30 musicians and compared it to the whole data to gain insight into how these top 30 artists' variable values differ from those of the other artists. To illustrate this comparison, we developed a radar chart. The chart above illustrates the various variable ranges of values between 0 to 0.8 for all the artists put together, including tempo, valence, energy, etc., by averaging them and displaying them on the appropriate radius for each respective variable.

- From the above visualization we can observe that the Top artists radar chart is infused upon all artists' values. If we look closely, we can find that the danceability and valence variables for the top listed musicians are nearly identical to those for all the other artists. This shows that the Top Artists Radar is infused upon all artists with respect to the elemental values. Therefore, we can say that for the most well-known artists, there isn't much of a difference or anything particularly special about this variable. The fact that the tempo variable is the same, as you can see, explains why most musicians favor the same tempo setting.
- However, if you compare the acousticness of the two, you'll find that the most well-known musicians have lower acousticness values. As a result, the acousticness of the visualization factors is crucial in the creation of popular music.

Visualization 9: Sunburst displaying Top 30 Artists w.r.t Genre

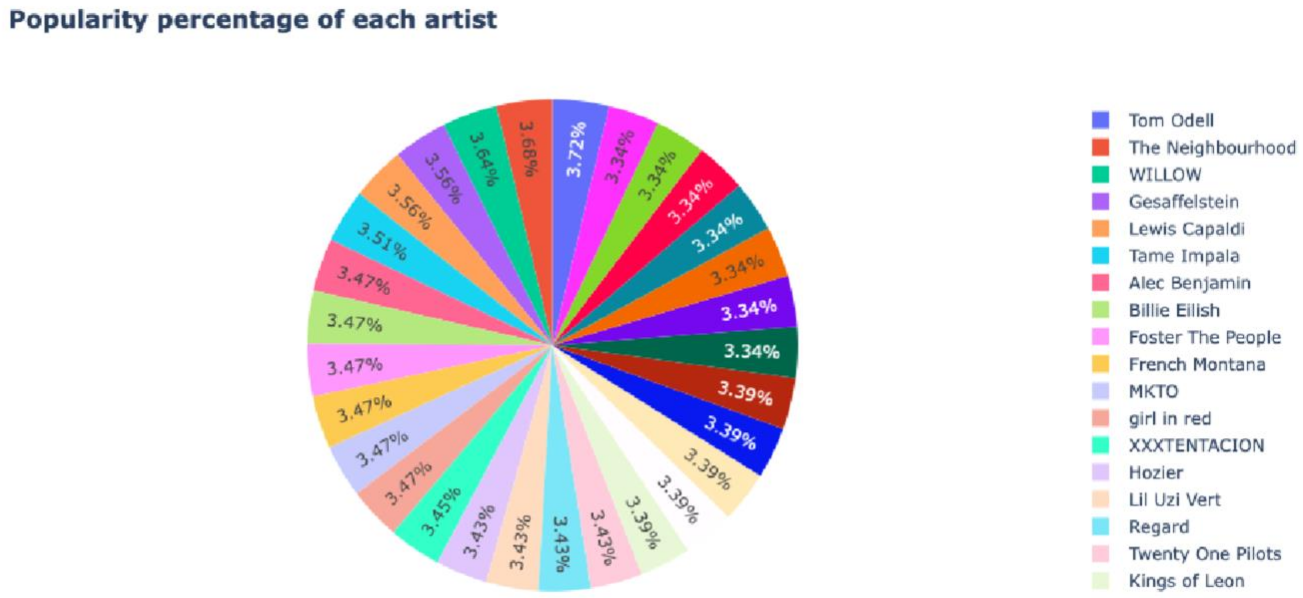


As explained earlier, genre plays a major role in any music; it is represented as the first classification method when analyzing music. Therefore, we wanted to analyze the genre for the top artists in a vision to predict the most popular genres chosen by top most artists alongside showcasing the different elemental variable values.

The above plot is called a sunburst plot and the reason why we have chosen this plot is because it categorizes and represents the artists based upon the genre as showcased in the visualization above.

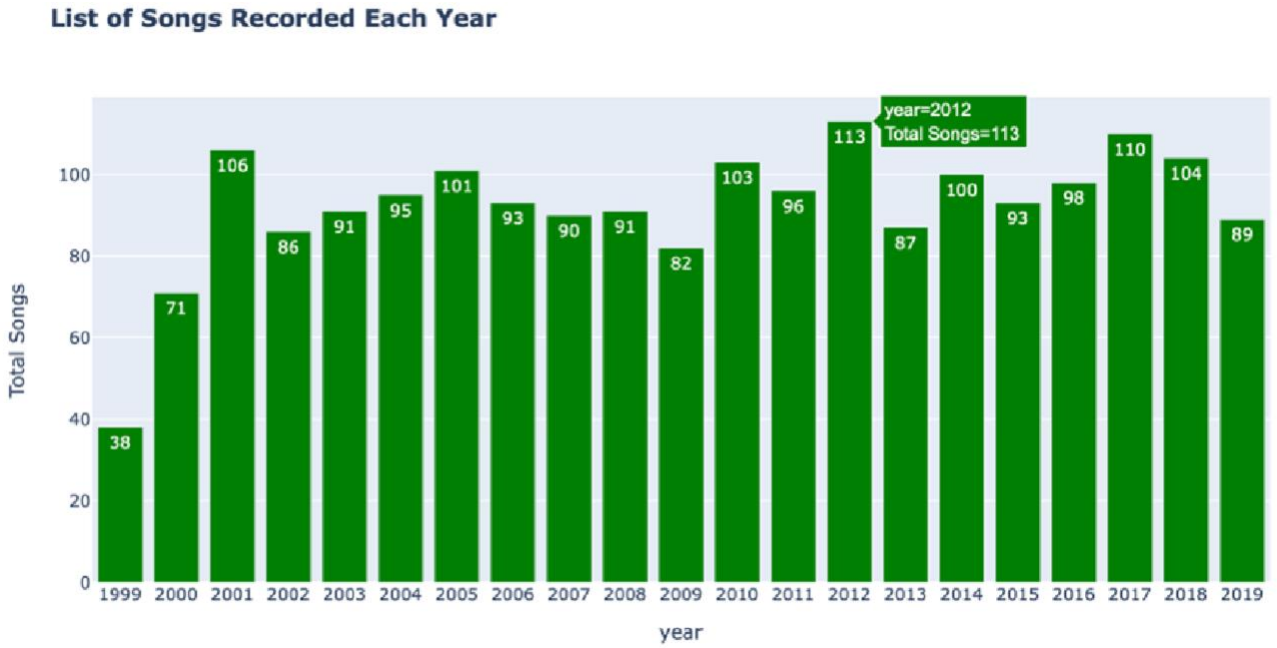
- From careful study, it is clear that pop music is the most popularly used genre by the most successful musicians, followed by rock music, hip-hop, and the rest. Plotly created the visualization, which highlights the associated values as you move your cursor over the data. The visualization is plotted using the plotly library and therefore showcases the values involved as you hover over the data.
- As we can see, Tom Odell, who selected the genre "pop," is the musician who received the highest ranking for popularity across the entire data set.

Visualization 10: Pie Chart



After examining the popularity rankings for the top 30 musicians, it appeared that they were in close proximity to one another. Which led us to question whether the popularity of each of the 30 artists on the list is distributed equally. In order to visualize the popularity estimated as a percentage overall, we designed a pie chart visualization that depicts the popularity percentage of each artist. The following image reveals that all of the artists in the top 30 have values that are almost identical to the 3.3% range for all of the percentages, explaining why no musician on the list is more well-liked than any other demonstrating that each musician on the top 30 list is almost equally popular.

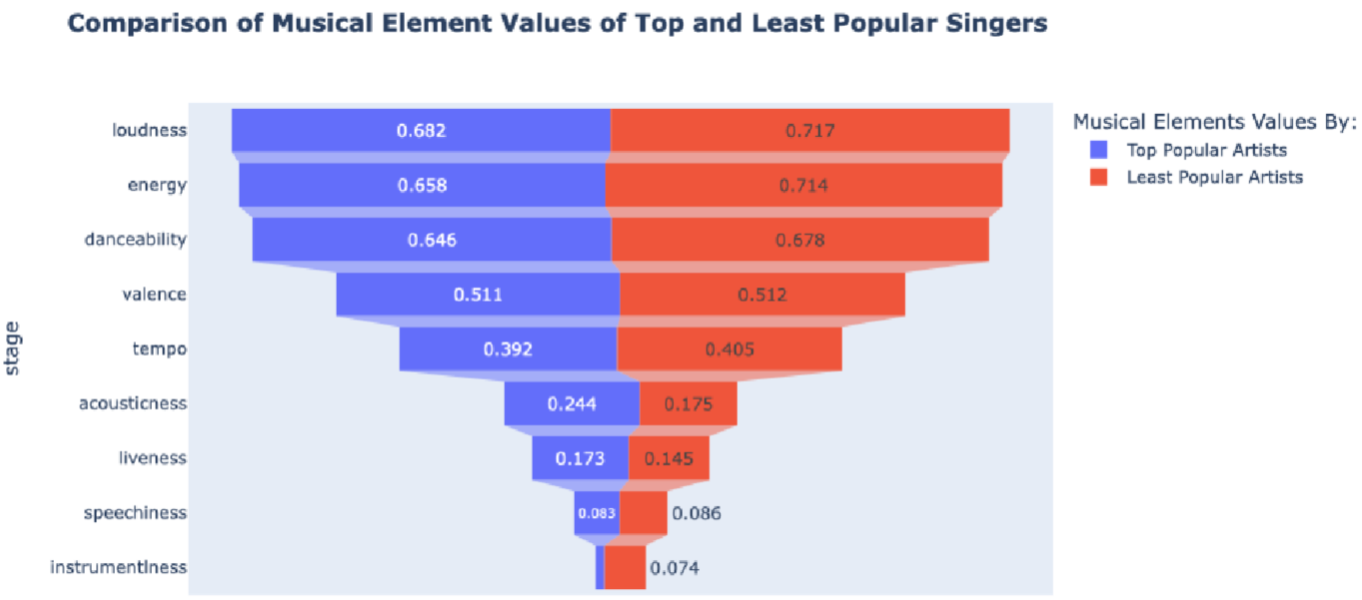
Visualization 11: Bar Chart



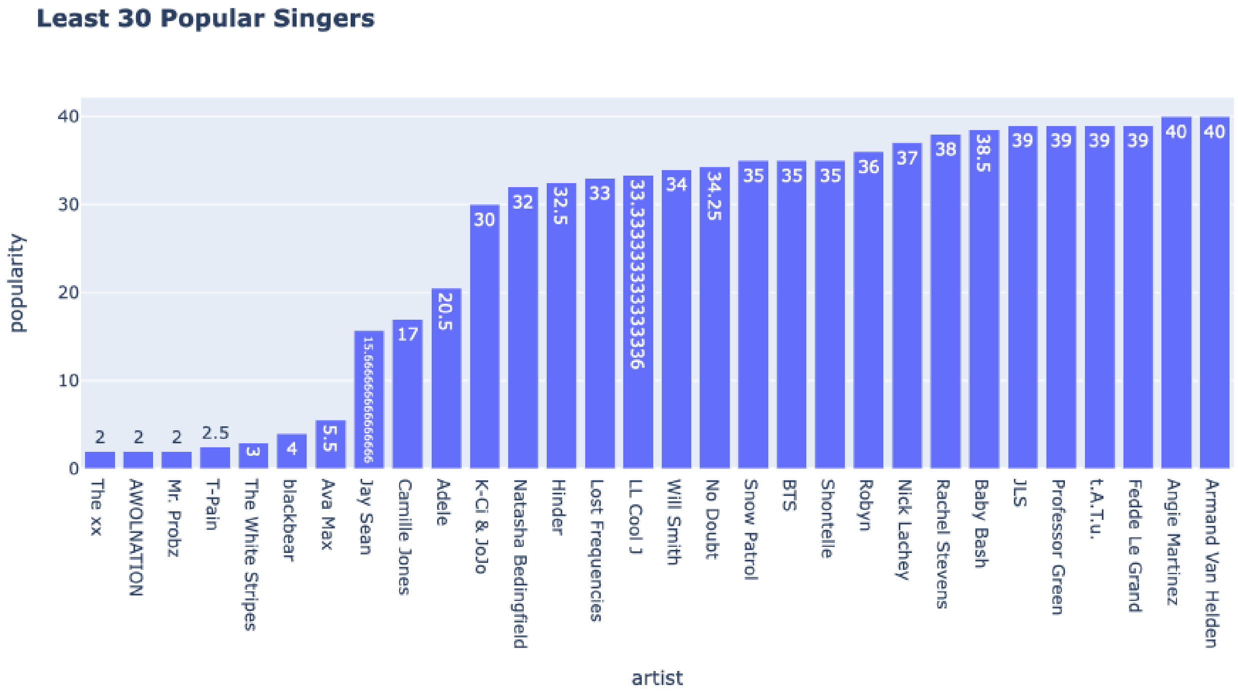
One of the most crucial findings in the insights that we sought to investigate was the total number of songs that appeared in the top list over the various years. The bar plot with the year versus. total song count was selected to illustrate this. When compared to the other years in the dataset, the year 2012 provided the most popular hits, as shown in the visualization above. However, there hasn't been a noticeable difference in the remaining years, which explains why there have been hits primarily every year. However, you'll notice that the number of song hits in 1999 and 2000 was quite low. This can be explained by the fact that there wasn't much technology available to showcase music all over the world in the late 19th century. However, with the technology available in the 20th century, an album can now be distributed to every corner of the world, explaining why there were no significant changes in the years of the 20th century. Furthermore, after careful analysis of the dataset, it was discovered that many well-known artists from the top 30 list, recorded their iconic albums after the year 2010, suggesting that perhaps this was the time when new musicians began to emerge.

Visualization 11: Funnel Chart to compare values taken by Top and Least popular artists

As already said, the primary goal of this project is to present the right musical element values for the fundamental variables that new and future musicians can utilize to produce music that will appeal to the broadest audience. But in order to demonstrate how important the variable values are, we plotted the visualization using the top 30 and bottom 30 artists from our dataset. By averaging the variable values across all 30 members of the two lists, we were able to better understand the differences and the significance of the variables and the values to be taken for popular hits.

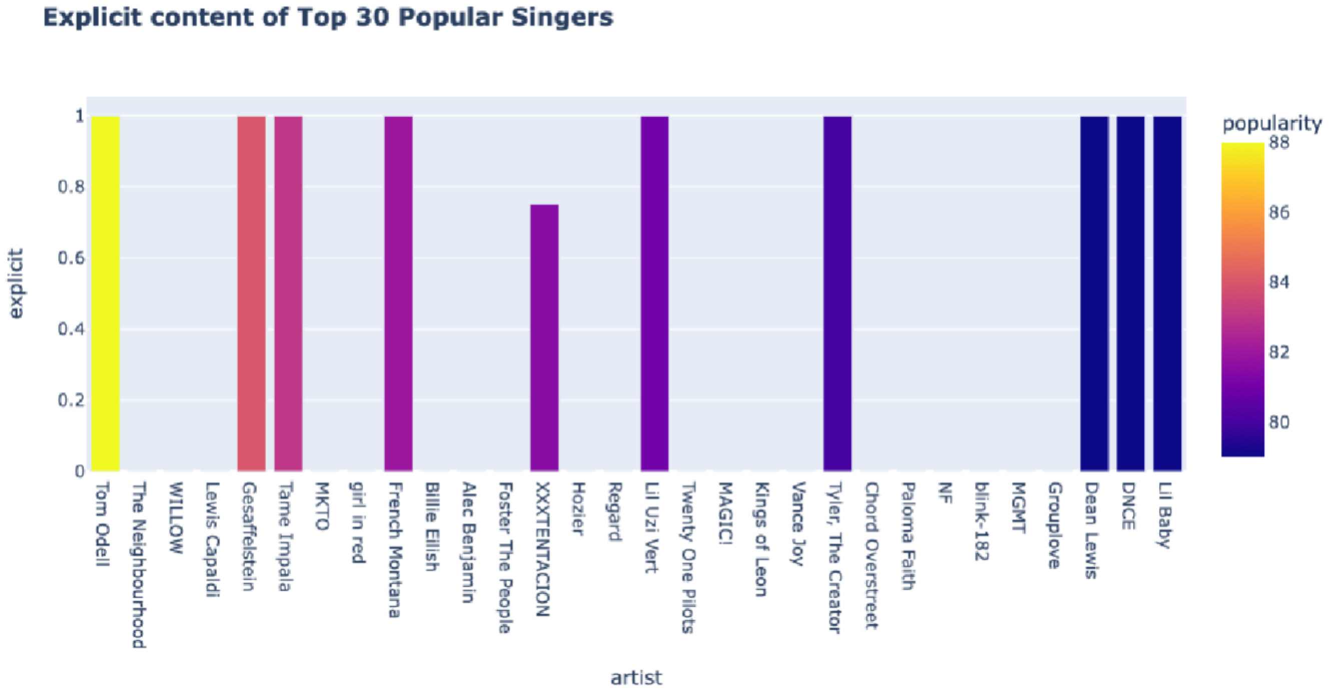


To compute the above funnel chart, we have taken top 30 artists list from the bar graph we mentioned above and the least 30 popular artists graph is shown below:



This graph is computed by averaging the popularity values and taking the bottom 30 artists from the dataset.

Visualization 12:Bar chart depicting explicit content



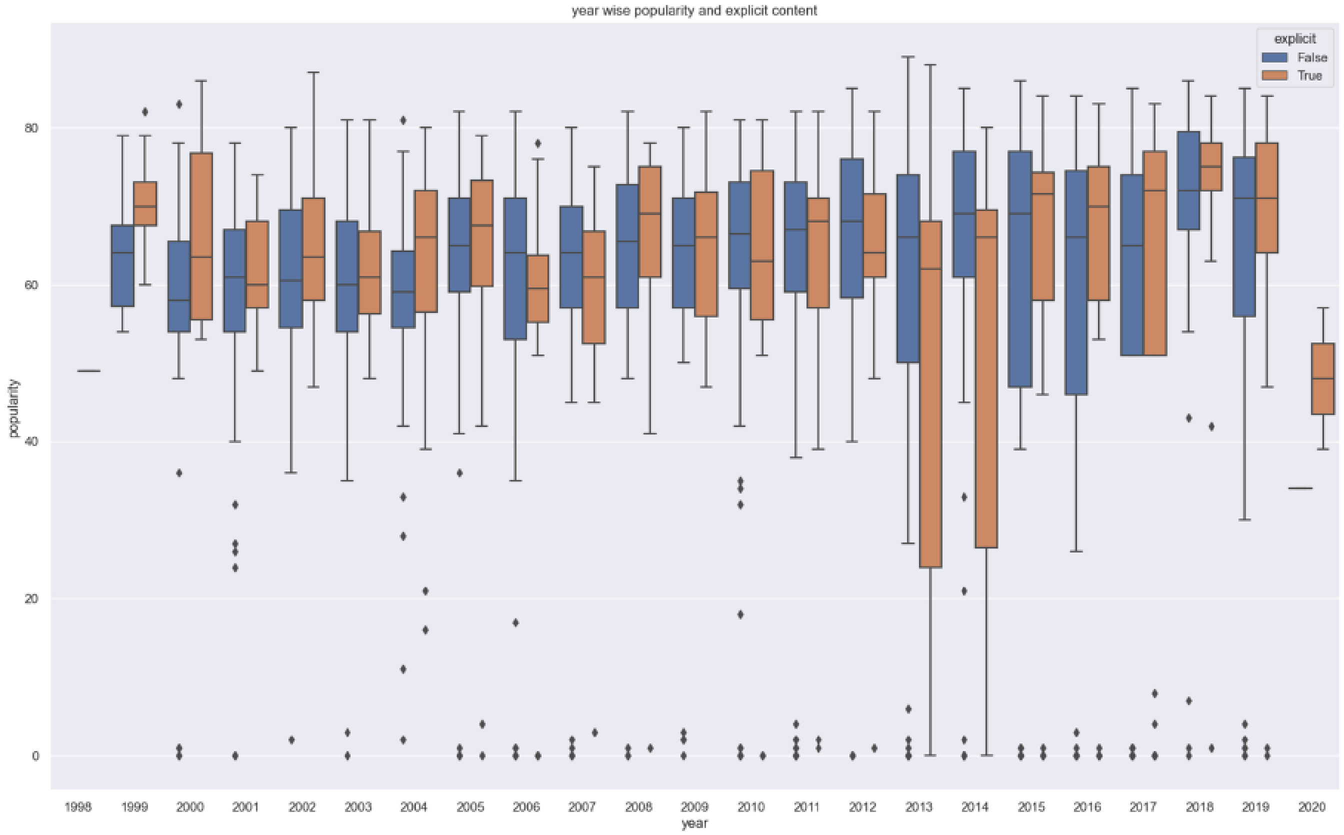
The explicit content in the dataset, which is the content of a song or a music which is in the lyrical content, is one of the significant factors that we have noticed when analyzing the dataset. This can be regarded as offensive or inappropriate for minors. As a result, we sought to evaluate this variable in order to predict whether or not popular songs include explicit content. Values 1 and 0 are used to signify the explicit content. As a result, we developed a visualization to determine whether the top 30 artists had used explicit content or not. However, to make the visualization even more intriguing, we also added the popularity value, which is shown above, so that the viewer could also learn about the popularity of the artist that is using explicit content.

Only 10 of the 30 singers in the graph above are using explicit content, indicating that most popular musicians refrain from using language that is inappropriate for children. Perhaps as a result, their music appeals to a wide audience and people of all ages can enjoy it. As many people listen to rap, which consists primarily of offensive words as we can see from the visualization, we can deduce that employing offensive words also helps you rise to the top of the list. Our favorite singer, XXXTENTACION, utilizes offensive language, but we still like his songs.

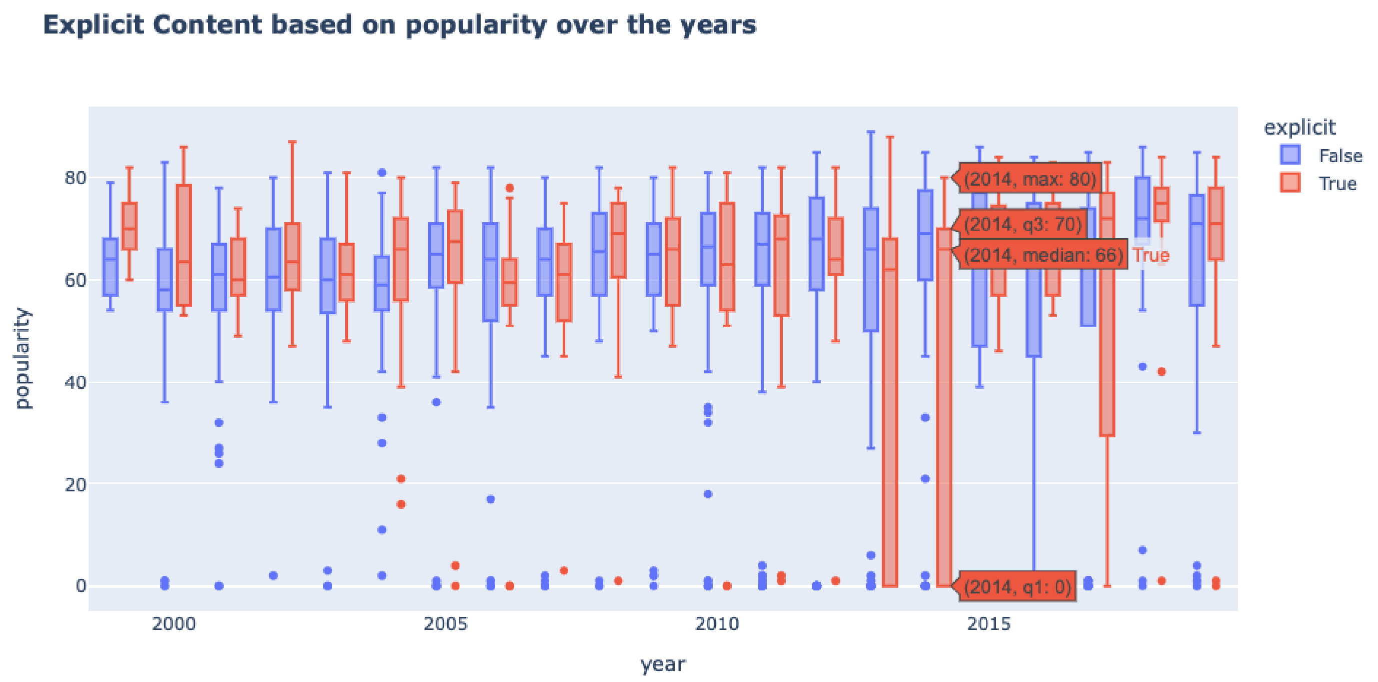
Visualization 13: Box Plot

To further understand about the explicit content in the data set we wanted to investigate it over the whole data set understanding its change over the years, in order to do so we have drawn a box plot depicting the relation between explicit content over the years with respect to popularity

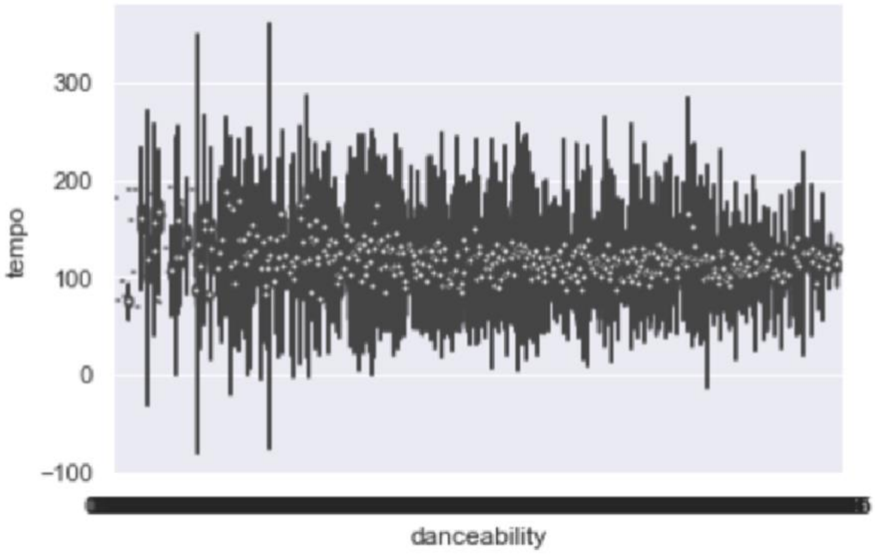
Below is an existing graph depicting the explicit content using box plot:



We wanted to further enhance this plot for effective visualization, so we produced an interactive visualization that would make it easier for people who are color blind to view the values when they are hovering over the data using a plotly library, as shown below. This makes it user readable and can also see the mean, median, and interquartile range values which helps us in understanding where most of the data points underlie.



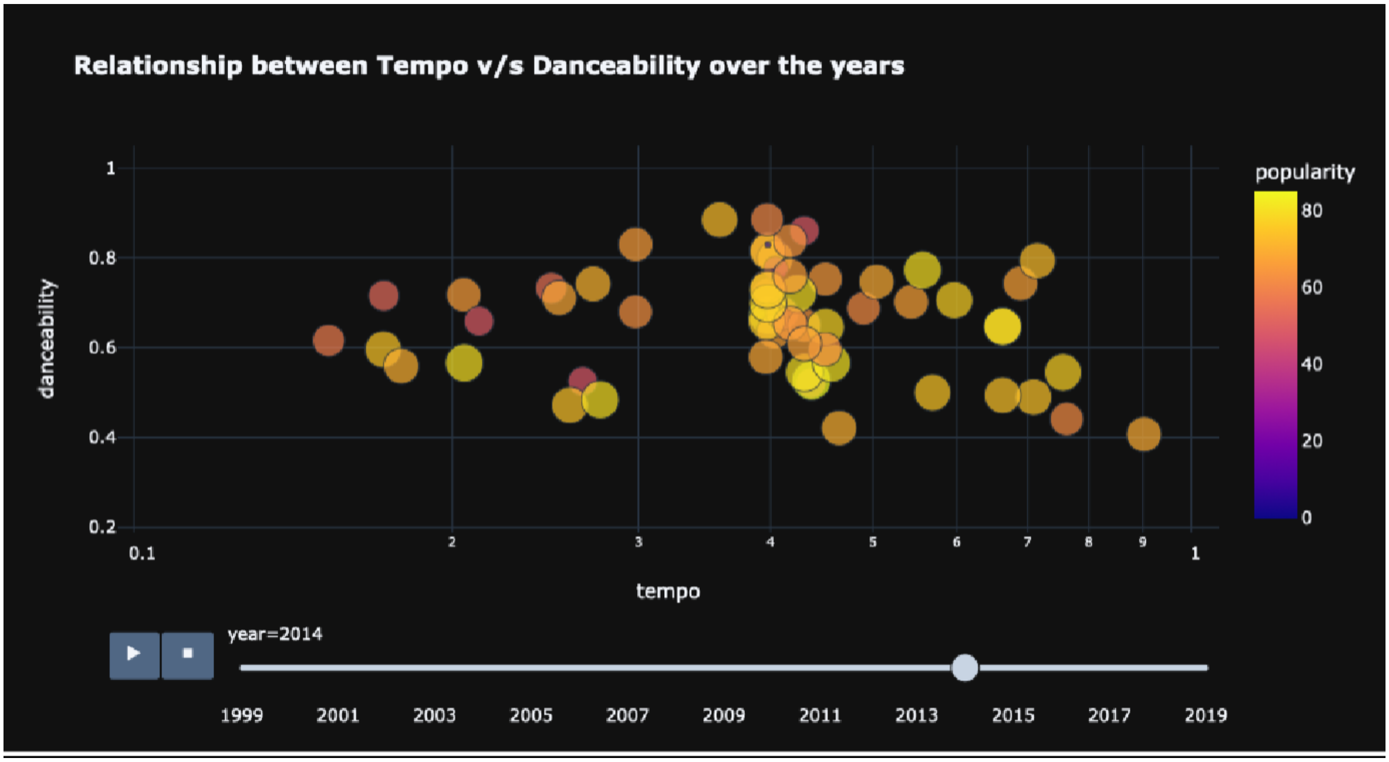
Visualization 14: Correlation of different variables



We wanted to find the relation between the different variables to understand or predict insights from it. But in the hunt to create a relational graph between the variables we have tried to plot the relation using a violin plot. The above graph depicts the relation between the tempo and danceability and from the above graph it's hard to predict the insights as the tempo values are infused inside the danceability and its hard to interpret the insights.

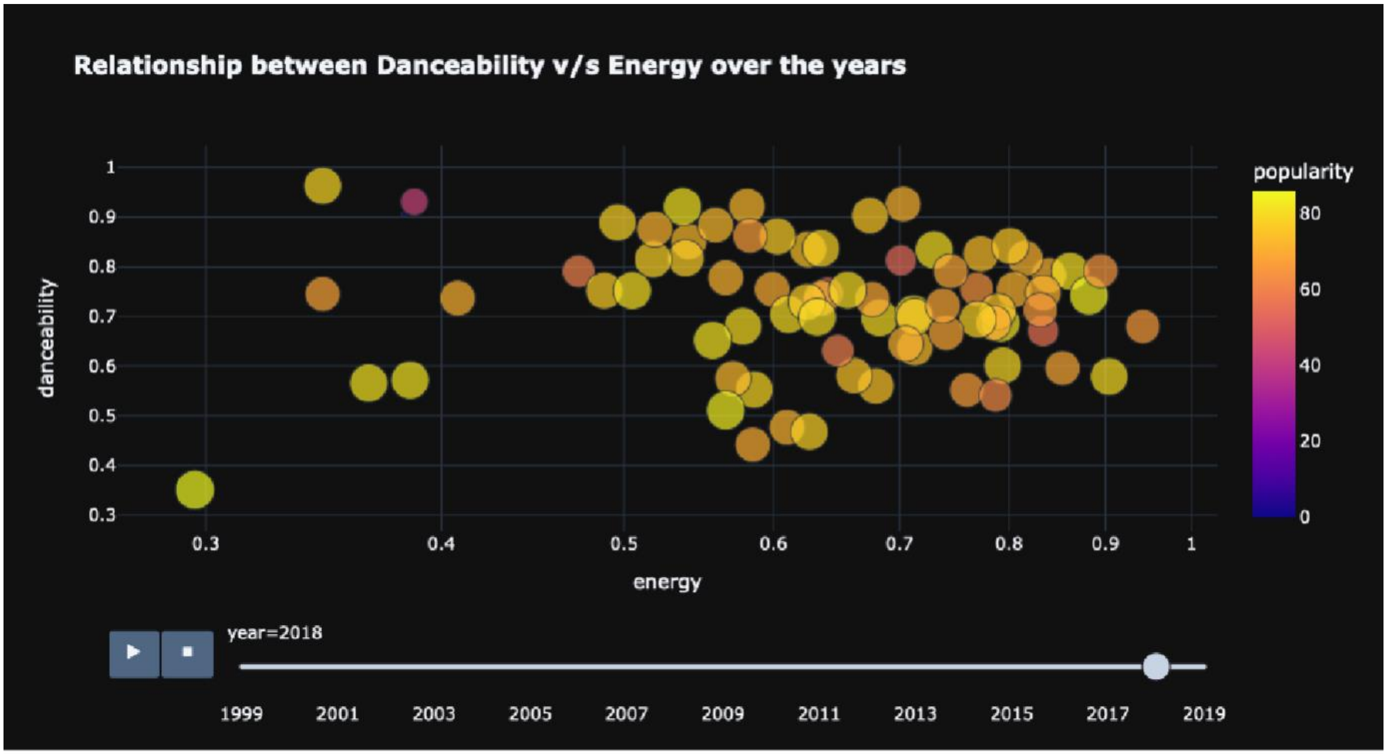
We looked at many variables and discovered the scatter plot that is already shown above in the existing data. We wanted to make it even better by removing the cluttered data points and make it more easily understandable and so we introduced years and determined the relationship between variables based on popularity metric as a moving visualization. This is the highlight of our project.

a)

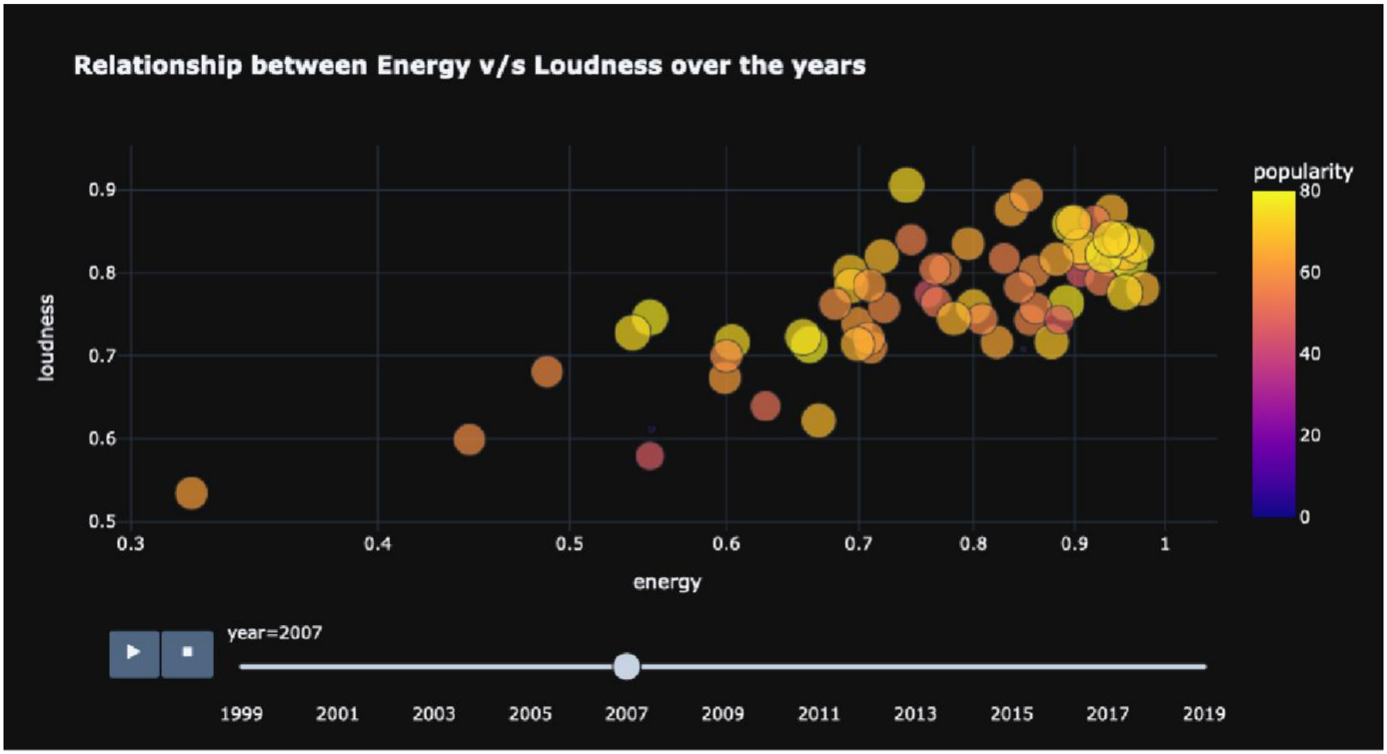


The relationship between tempo and danceability over time is shown in the above animated image. Since the data is spread out over many years, clustering is also avoided because the data points are simple to understand. From the graph above, we can see that from 2011 to 2015, the tempo data is more consistently at a value of 0.4, however for the remaining years, the data is dispersed. The higher danceability and tempo values are more popular as observed from the overall years.

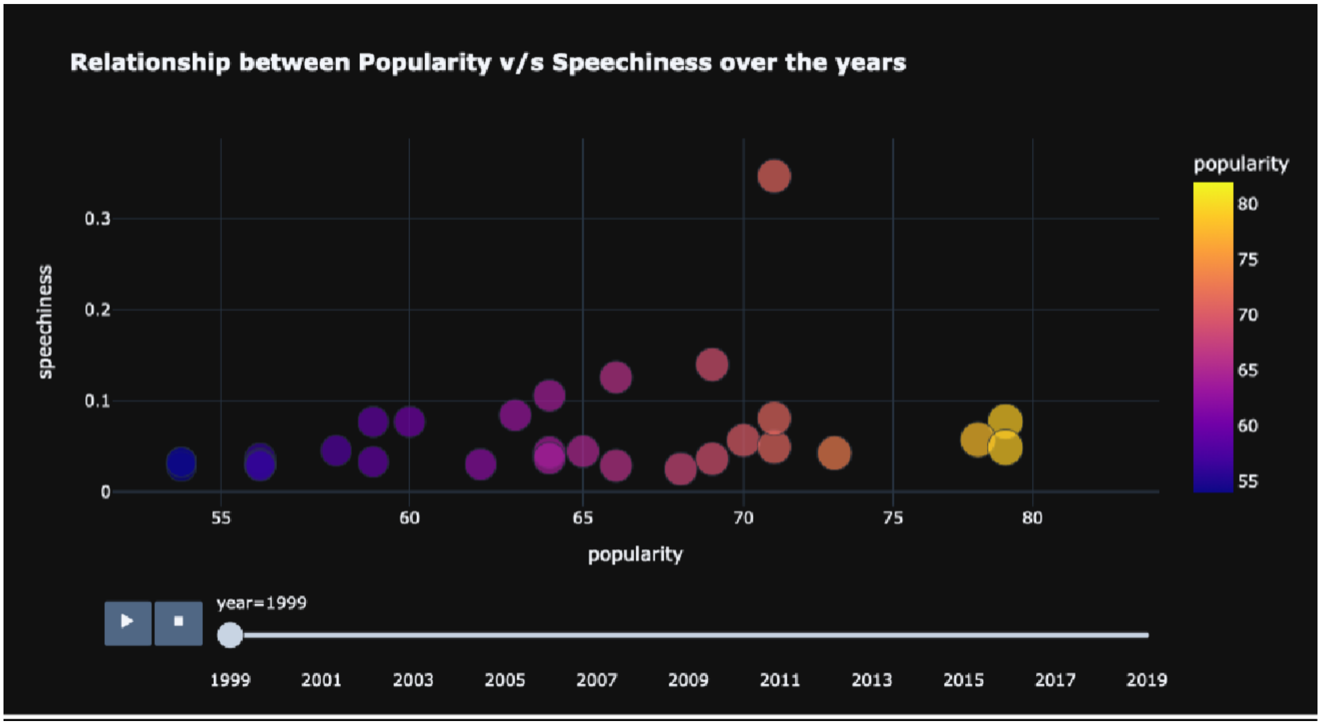
b) The second graph depicts the relationship between danceability and energy. It's observed that higher the danceability and energy values, the more popular the song is. Almost all the popular songs have higher danceability value but for most of the songs have energy values between 0.69 to 0.9. In 2018, almost all the songs are popular and have combinations of values in the range of 0.6 to 0.8 for danceability and 0.55 to 0.85 energy. So, maybe the combinations used in 2018 by artists could be used to produce musical hits.



c) The below energy and loudness relationship is closely observed, we can infer from this graph that over the years, most of the popular songs have values for energy in the range of 0.7 to 0.9 and loudness in the range of 0.8 to 0.9. But in 2018, almost all the values ranging between 0.6 to 0.8 are highly popular. This could be another range of values to be considered for composing good music.

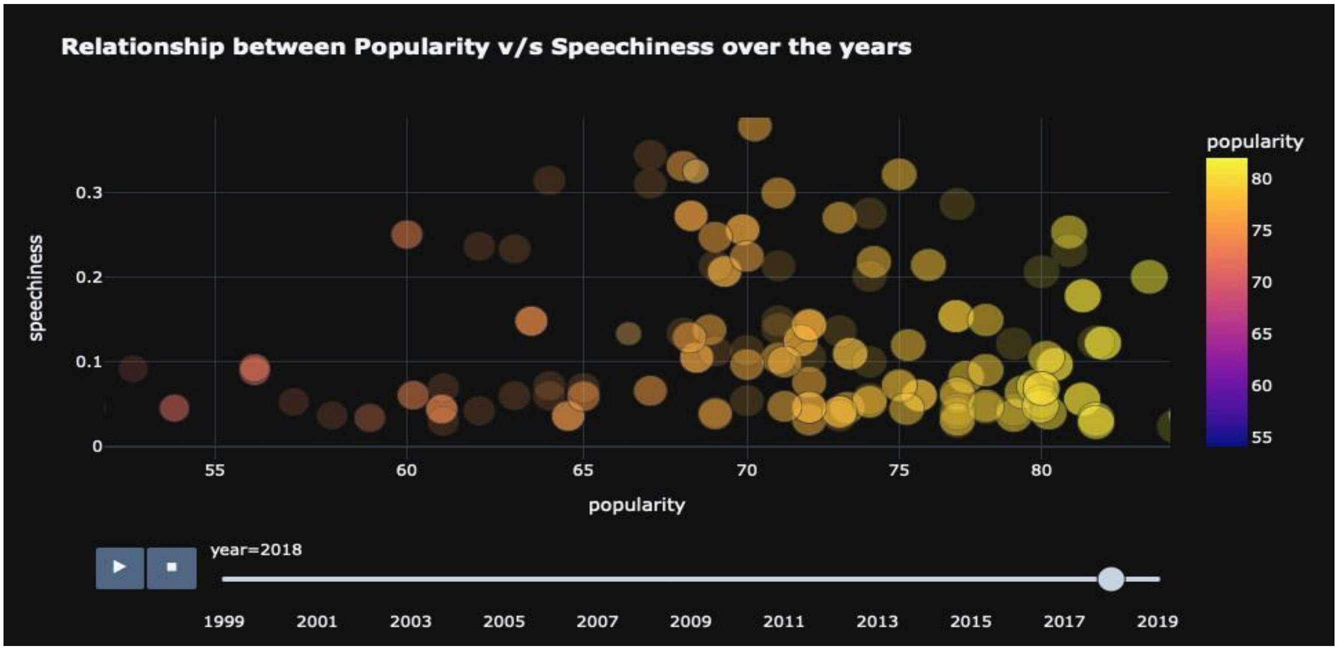


d)



The association between speechiness and popularity over time is depicted in the graph up top. It is challenging to extrapolate any conclusions from 1999 because the data is dispersed across the year. We may play the button on this graph, which also displays the trends in the data points as the year goes by. From the overall popularity trend, it is clear that songs with lower speechiness scores tend to be more popular, but in the years 2007, 2009, 2011, 2017, and 2019, it was shown that songs with greater speechiness values tend to be more well-liked. Due to the equal popularity of songs in both categories in recent years, it is difficult to determine the ratio of speechiness songs to non-speechiness songs.

In Recent years:



DISCUSSION & CONCLUSION:

1. As we have observed that there is a difference in the variable values between the least popular and most popular artists, we can conclude using these depicted most popular artist's elemental variable values depicted through the funnel graph that can help a new artist compose a heart winning music.
2. We represented the analysis for the different variables using various visualizations in relation to popularity, and we came to the conclusion that some fundamental variable components, such as the tempo, have remained the same while some featuring variables, such as the acousticness, danceability, energy, loudness, explicit content play a significant role. Additionally, listers favor the pop genre more and this genre has high danceability and energy values compared to other genres which explains the popularity.
3. The bar graph shown above shows the number of songs played each year, and we can see that 2012 has the most hits since, according to our data, many new artists had their biggest hits in 2011 and because of modern technology, the number of hits is strongly correlated between years and is maintaining closely across recent years.
4. We have portrayed many visualizations in respect to the genre, which is important, and we have discovered that the pop genre is more popular.
5. As discussed earlier we wanted to analyze the relation between different variables by plotting an efficient graph. Therefore, we have created the moving scatter plot visualization with respect to the years and popularity reducing the inefficiency of the clustering of the data points.
6. Using the plots treemap and sunburst, we created several effective and eye-catching visualizations to highlight the various values and relationships between the variables as displayed above which any viewer can easily understand.
7. We have investigated the explicit content for both the top artists and the whole dataset as showcased above in the results section depicting insights such as that the most popular songs have less explicit content ratio.

LIMITATIONS:

We can first draw the conclusion that there is unquestionably a difference in the variable values for the popular artists when compared to the least popular artists, and that these visualizations serve as an explanation and an answer showcasing the numerically correct values that are used by those artists.As previously indicated, any aspiring musician aiming to create music that captures listeners' hearts could use these visuals as a guide and work with these metrics to succeed. The dataset we have chosen is smaller as we have only analyzed top hits from Spotify, thus we cannot predict for a huge base of artists and when the dataset is larger, we might discover new trends and patterns as opposed to what we have predicted.

As we have seen over the years, many new musicians have emerged, which has led us to believe that the music industry is changing swiftly. These values can be used as a guide when writing good music, but they cannot guarantee or confirm the success of an album as there are many other factors that must come into play these days for a musical album to be a success. The trend of music changes from year to year, and this cannot be predicted. Any analysis that we perform is only gonna be useful for current use and there are many more elements to be considered for making good music.

FUTURE WORK:

Because of the limitations mentioned above, it is difficult to show exact values for the majority of popular music because music trends fluctuate from year to year. In order to comprehend the causes of the change and to forecast various insights that aid in anticipating the next change, we therefore want to focus our future work on the music trend. Working on the music trend, in our opinion, is crucial since it facilitates understanding listeners' perspectives, which is the first step in creating the best music.

REFERENCES:

1) https://www.kaggle.com/code/goodwanghan/spotify-data-eda-with-fugue-sql-duckdb-and-dask/notebook
2) https://www.kaggle.com/code/shivagowri19/spotifyanalysis/notebook
3) https://www.kaggle.com/code/tanakalusengo/spotify-popularity-classification-models/notebook
4) https://plotly.com/python/px-arguments/