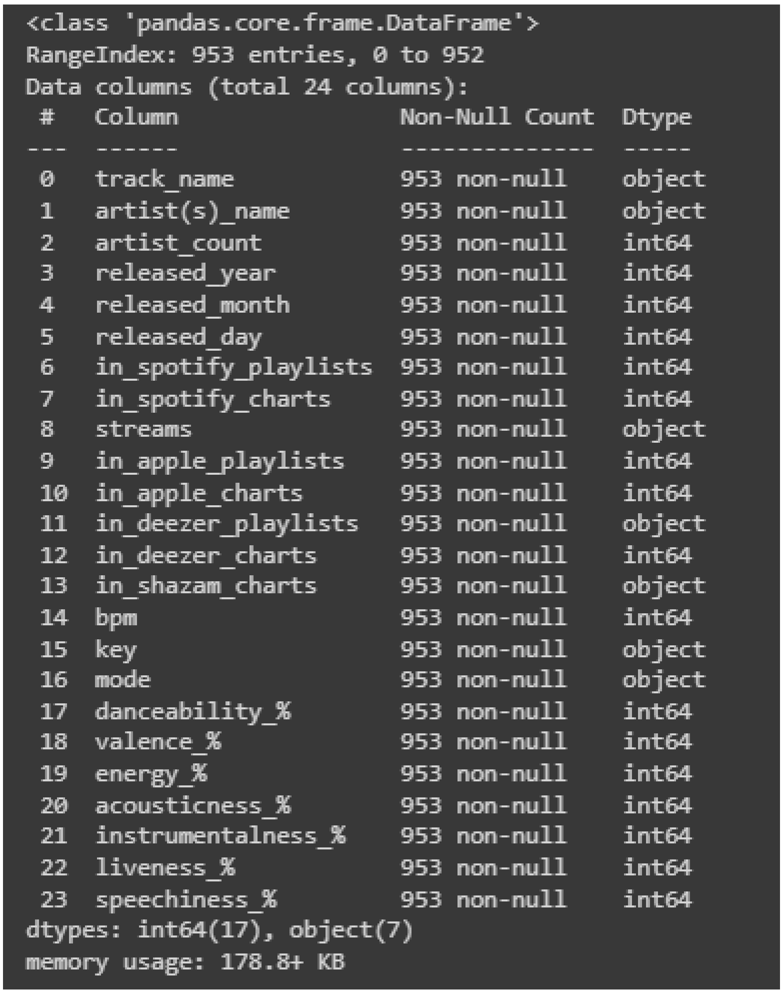
Honors Peer-graded Assignment

Exploratory Data Analysis

Michał Bożyk

1. Dataset summary

The dataset that was chosen concerns most streamed songs on Spotify in year 2023 (https://www.kaggle.com/datasets/nelgiriyewithana/top-spotify-songs-2023?resource=download). It contains 953 rows and 24 columns, which have labels such as artist names, bpm (beats per minute) or the number of streams. Dataset also contains information on danceability, energy or instrumentalness which are expressed in percentages.



```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 953 entries, 0 to 952
Data columns (total 24 columns):
 #   Column                Non-Null Count  Dtype 
---  ------                --------------  ----- 
 0   track_name            953 non-null    object
 1   artist(s)_name        953 non-null    object
 2   artist_count          953 non-null    int64 
 3   released_year         953 non-null    int64 
 4   released_month        953 non-null    int64 
 5   released_day          953 non-null    int64 
 6   in_spotify_playlists  953 non-null    int64 
 7   in_spotify_charts     953 non-null    int64 
 8   streams               953 non-null    object
 9   in_apple_playlists    953 non-null    int64 
 10  in_apple_charts       953 non-null    int64 
 11  in_deezer_playlists   953 non-null    object
 12  in_deezer_charts      953 non-null    int64 
 13  in_shazam_charts      953 non-null    object
 14  bpm                   953 non-null    int64 
 15  key                   953 non-null    object
 16  mode                  953 non-null    object
 17  danceability_%        953 non-null    int64 
 18  valence_%             953 non-null    int64 
 19  energy_%              953 non-null    int64 
 20  acousticness_%        953 non-null    int64 
 21  instrumentalness_%    953 non-null    int64 
 22  liveness_%            953 non-null    int64 
 23  speechiness_%         953 non-null    int64 
dtypes: int64(17), object(7)
memory usage: 178.8+ KB
```

2. Plan for data exploration

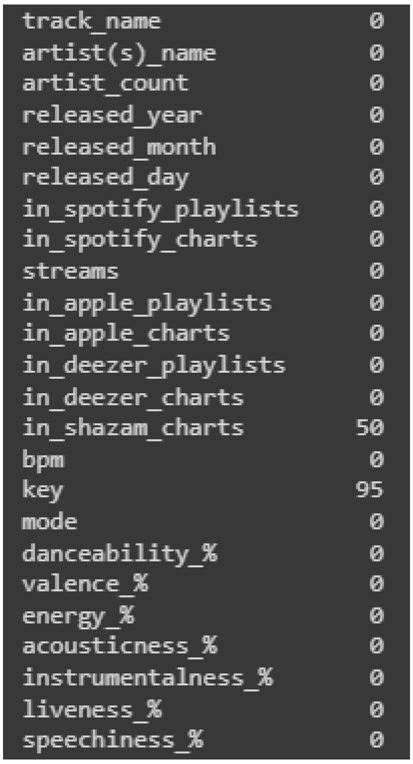
First data cleansing and feature engineering will be conducted. After that the following analyses will be performed:

a) check the top 10 most streamed songs,
b) check which artist is the most common on the list and what years the songs are from,
c) check which songs that came out in 2023 had the fastest growth in number of streams,
d) check the correlation between bpm and different audio features,
e) determine how audio features were changing over the years.

Next, hypotheses about the data will be formulated and one of them will be tested. The last part of the assignment will concern suggestions for further analysis and the summary of the overall quality of the dataset.

3. Data cleansing and feature engineering

First the check for NaN values was performed. The search yielded the following results:

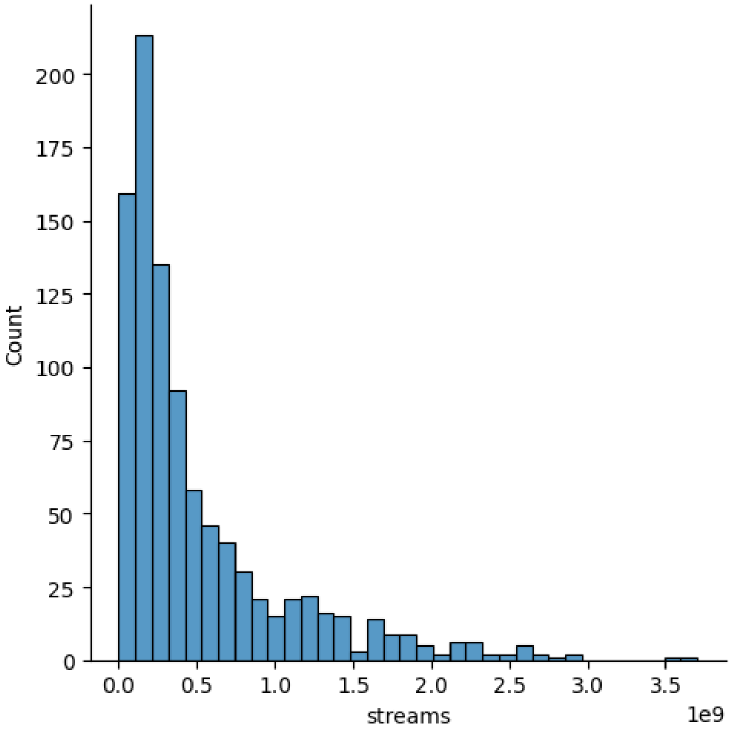


```
track_name              0
artist(s)_name          0
artist_count            0
released_year           0
released_month          0
released_day            0
in_spotify_playlists    0
in_spotify_charts       0
streams                 0
in_apple_playlists      0
in_apple_charts         0
in_deezer_playlists     0
in_deezer_charts        0
in_shazam_charts       50
bpm                     0
key                    95
mode                    0
danceability_%          0
valence_%               0
energy_%                0
acousticness_%          0
instrumentalness_%      0
liveness_%              0
speechiness_%           0
```

Null values were replaced with forward fill method.

Since, the column with label "streams" (number of streams) is arguably the most important it will be converted to numeric data to examine its correlations with other features. However, it turned out that one of the rows contained type the could not be converted to numeric data, therefore median value was assigned to this data.

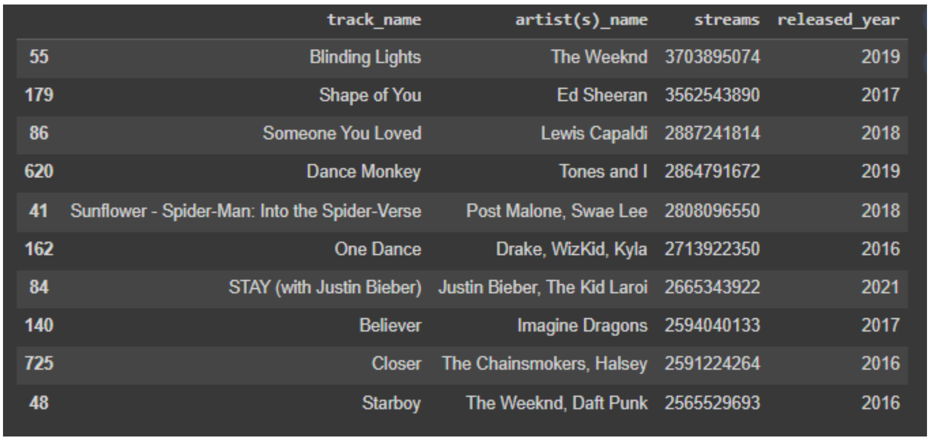
It was examined whether the “streams” data is normally distributed. It turned out that it is not, and is relatively skewed (the skewness is equal to 2).



Further analysis concerned checking whether, duplicates occur in the dataset, however, it yielded no results. There are also no significant outliers that may spoil the overall analysis.

4. Key Findings and Insights

First, it was found out which 10 songs had the most streams in 2023.

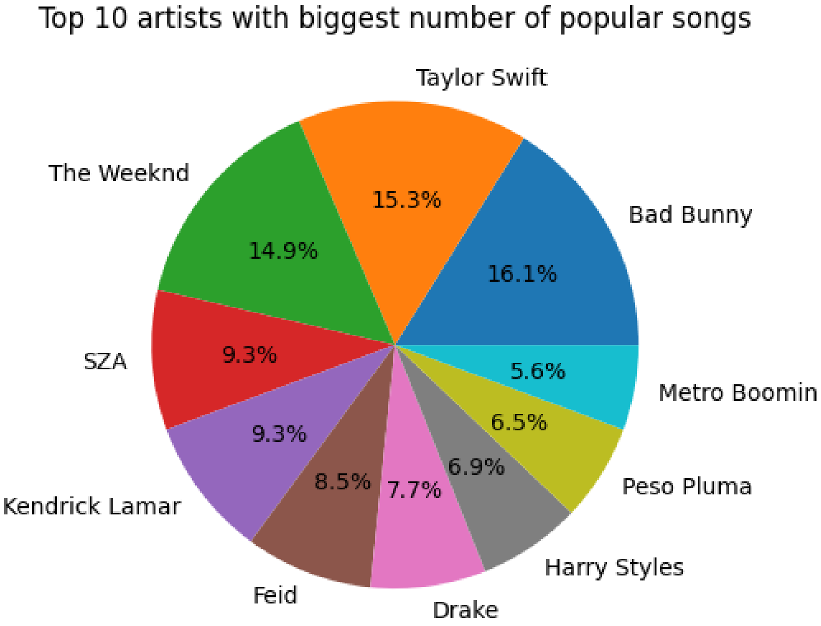


	track_name	artist(s)_name	streams	released_year
55	Blinding Lights	The Weeknd	3703895074	2019
179	Shape of You	Ed Sheeran	3562543890	2017
86	Someone You Loved	Lewis Capaldi	2887241814	2018
620	Dance Monkey	Tones and I	2864791672	2019
41	Sunflower - Spider-Man: Into the Spider-Verse	Post Malone, Swae Lee	2808096550	2018
162	One Dance	Drake, WizKid, Kyla	2713922350	2016
84	STAY (with Justin Bieber)	Justin Bieber, The Kid Laroi	2665343922	2021
140	Believer	Imagine Dragons	2594040133	2017
725	Closer	The Chainsmokers, Halsey	2591224264	2016
48	Starboy	The Weeknd, Daft Punk	2565529693	2016

As expected songs with most streams are couple of years old. The newest one was released in 2021.

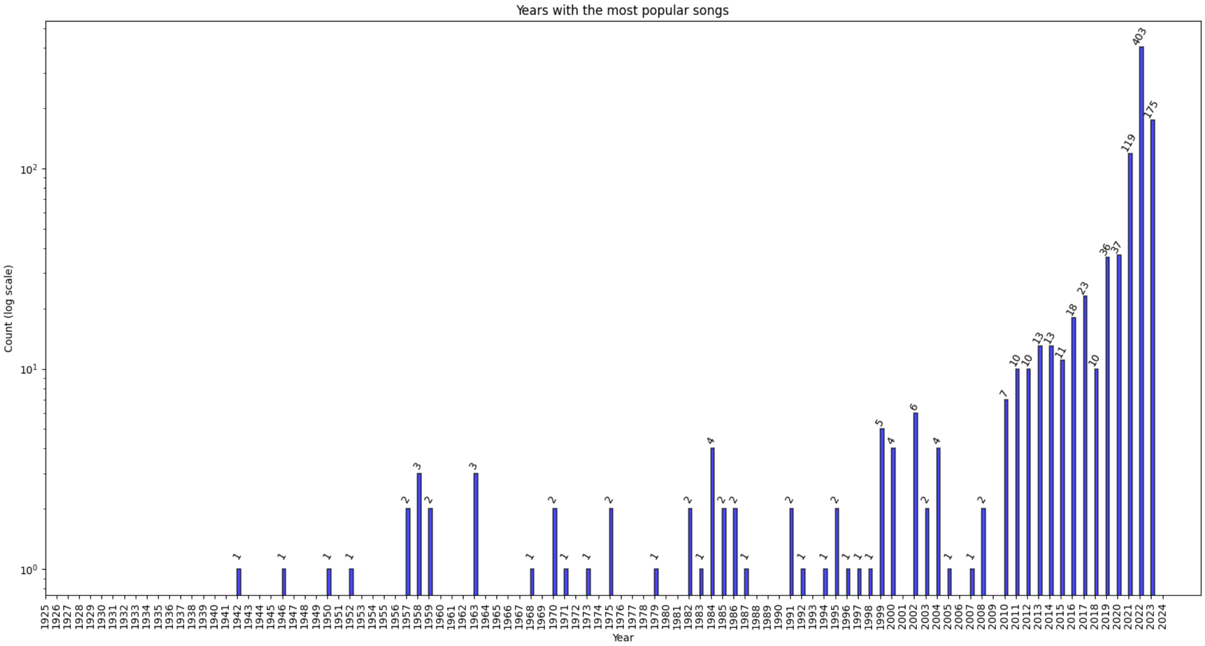
Next, checked was the most common artist on the list as well as the years most of the songs are from.

Because of the probability that more than artist is involved in one song, artists name were separated based on comma. The results concerning top 10 artists with the biggest number of popular songs were shown on pie chart.

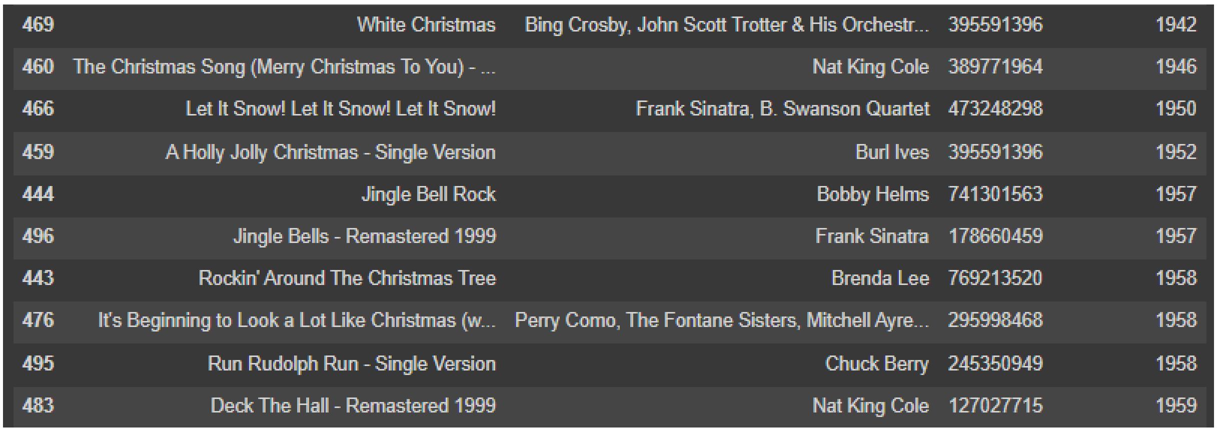


Top 3 artists (Bad Bunny, Taylor Swift and The Weeknd) constituted closely to 50%.

Next, the analysis regarding the year with the most popular songs was conducted. To better visualize the trend bar plot was generated. The y-axis (count of popular songs released in each year) was normalized using logarithmic scale.

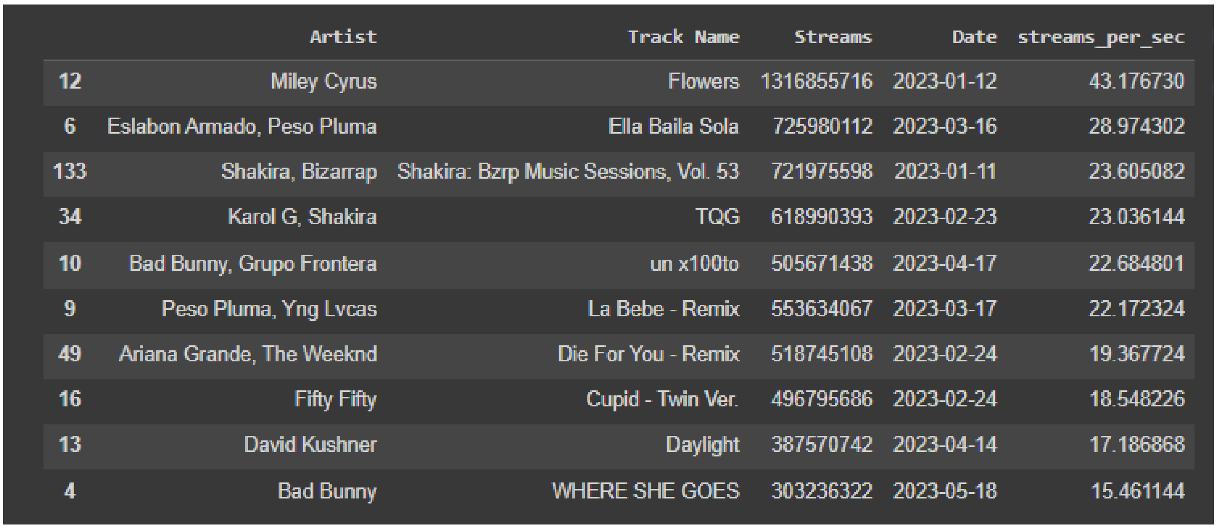


As it turned out, most of the most streamed songs (c. 73%) were released within past 2 years (2021, 2022, 2023). As for the oldest songs on the list spotted was one mistake concerning a song supposedly being released in 1930. The right date was corrected based on external sources. Interestingly, 10 oldest songs on the list are related to Christmas.



469	White Christmas	Bing Crosby, John Scott Trotter & His Orchestr...	395591396	1942
460	The Christmas Song (Merry Christmas To You) - ...	Nat King Cole	389771964	1946
466	Let It Snow! Let It Snow! Let It Snow!	Frank Sinatra, B. Swanson Quartet	473248298	1950
459	A Holly Jolly Christmas - Single Version	Burl Ives	395591396	1952
444	Jingle Bell Rock	Bobby Helms	741301563	1957
496	Jingle Bells - Remastered 1999	Frank Sinatra	178660459	1957
443	Rockin' Around The Christmas Tree	Brenda Lee	769213520	1958
476	It's Beginning to Look a Lot Like Christmas (w...	Perry Como, The Fontane Sisters, Mitchell Ayre...	295998468	1958
495	Run Rudolph Run - Single Version	Chuck Berry	245350949	1958
483	Deck The Hall - Remastered 1999	Nat King Cole	127027715	1959

Further analysis concerned checking which songs released in 2023 had the most streams per second since its release. It provides insights on which songs from 2023 had the highest increase of popularity. To assess it, three columns (released_year, released_month and released_year) were merged together and converted to datetime format. The date was later subtracted from 2023-12-31 (no information on the last day of gathering data was found, so the last day of the year was assumed). The results are shown below.

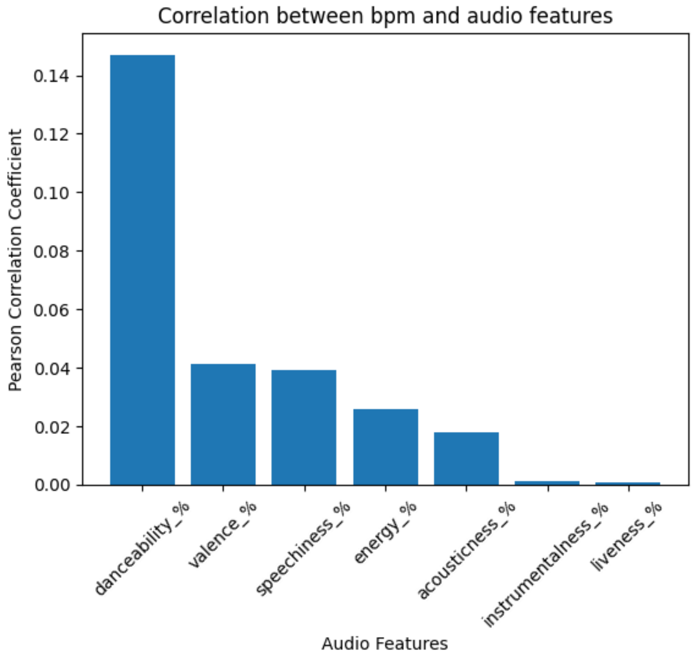


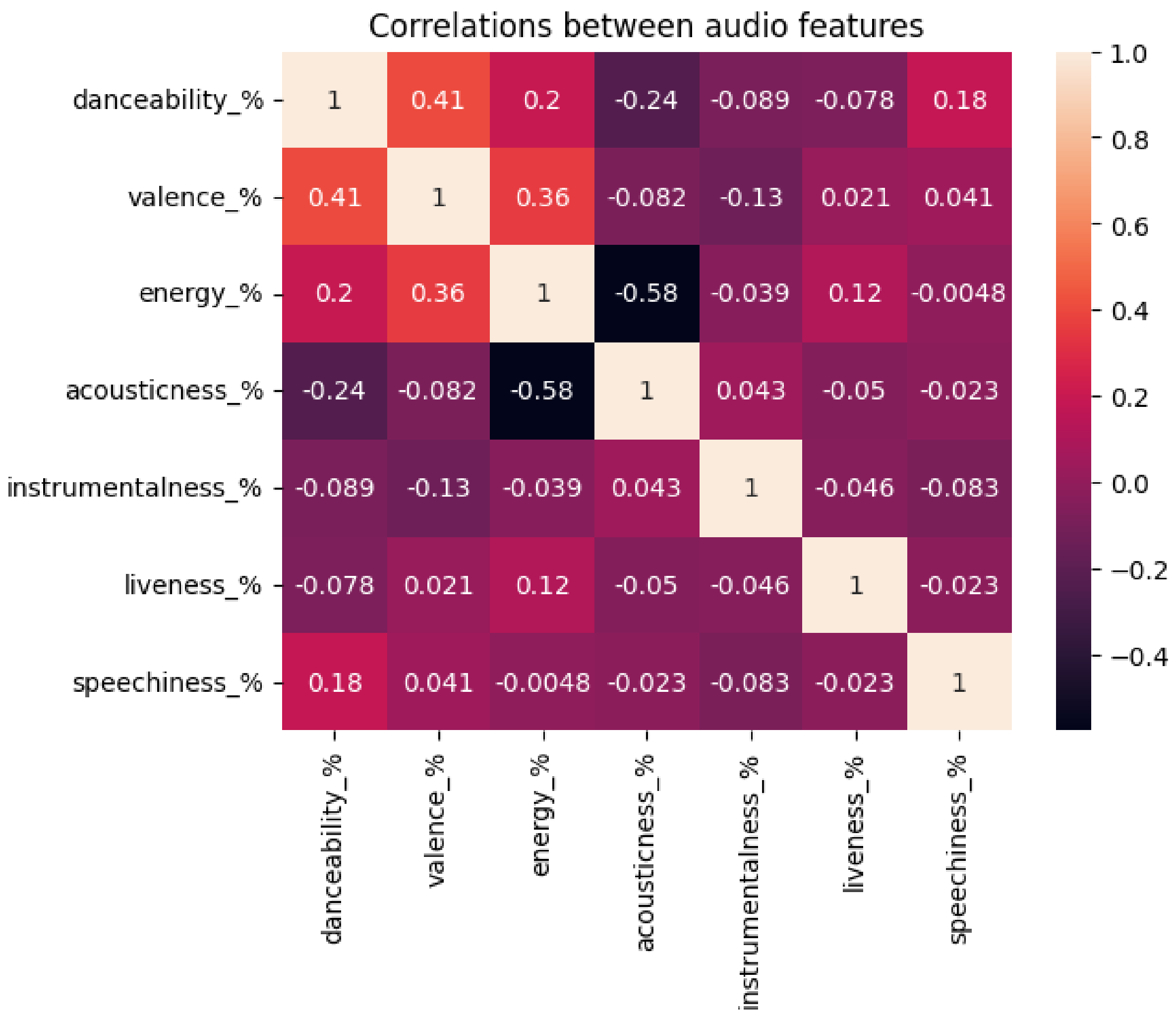
	Artist	Track Name	Streams	Date	streams_per_sec
12	Miley Cyrus	Flowers	1316855716	2023-01-12	43.176730
6	Eslabon Armado, Peso Pluma	Ella Baila Sola	725980112	2023-03-16	28.974302
133	Shakira, Bizarrap	Shakira: Bzrp Music Sessions, Vol. 53	721975598	2023-01-11	23.605082
34	Karol G, Shakira	TQG	618990393	2023-02-23	23.036144
10	Bad Bunny, Grupo Frontera	un x100to	505671438	2023-04-17	22.684801
9	Peso Pluma, Yng Lvcas	La Bebe - Remix	553634067	2023-03-17	22.172324
49	Ariana Grande, The Weeknd	Die For You - Remix	518745108	2023-02-24	19.367724
16	Fifty Fifty	Cupid - Twin Ver.	496795686	2023-02-24	18.548226
13	David Kushner	Daylight	387570742	2023-04-14	17.186868
4	Bad Bunny	WHERE SHE GOES	303236322	2023-05-18	15.461144

As for the further analysis, the correlation between bpm [beats per minute] and provided audio features were checked. Beats per minute give information on the tempo of musical composition. Audio features in the dataset are:

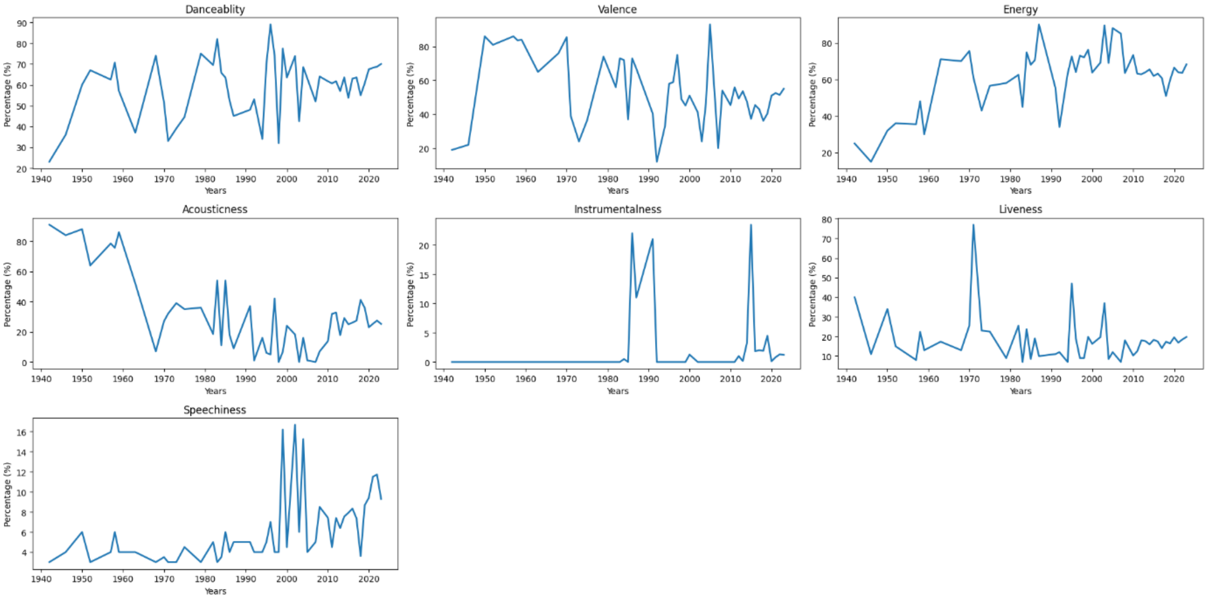
- danceability – how suitable the track is for dancing,
- valence – positiveness of the track,
- energy – intensity of the track,
- acousticness – how acoustic is the track,
- instrumentalness – whether the track is vocal or not,
- liveness – whether there is an audience recorded on the track,
- speechiness – whether there are spoken words in the track.

All of them are provided in percentages.





The last part of the analysis concerned the change of audio features over the years. Mean value of every audio feature in every year was calculated and 7 relevant plots were generated.



Plots show overall trends in the change of certain audio features. It is, of course, limited to the songs present in dataset, so it’s possible that results may differ, if more songs would have been taken into consideration. Plots show that, older songs tend to be more acoustic compared to modern ones. New songs, however, are more energetic than older ones.

Summary - key findings and insights:

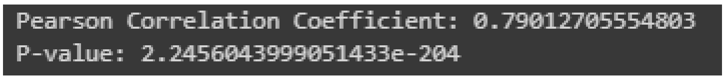
- top 10 most streamed songs are couple of years old with newest being from 2021,
- modern songs were the most popular, whereas 10 oldest songs were related to Christmas,
- song released in 2023 with the fastest growth of popularity had an average of c. 43 streams per second,
- beats per minute have weak correlation with audio features (the strongest correlation with danceability with coefficient of 0.14),
- the strongest positive correlation is between valence and danceability and valence and energy, whereas the strongest negative correlation is between acousticness and energy meaning if the songs is more acoustic it is also less energetic,
- there was visible decrease in acousticness over the years. Modern songs however tend to be more energetic than the older ones.

5. Formulating hypotheses and testing

3 hypotheses were formulated:

1) tracks that occur in higher number of Spotify charts and playlists accumulated more streams,
2) tracks present in Apple playlists have less streams comparing to those only included on Spotify playlists,
3) collaborations between artists tend to gain more streams.

Chosen was 1) hypothesis. Conducted was Pearson correlation test. The results are presented below:



```
Pearson Correlation Coefficient: 0.79012705554803
P-value: 2.2456043999051433e-204
```

Considering alpha being 0.05 P-value is below that number therefore the test is statistically significant. Moreover, there is positive correlation between the number of Spotify playlists that include songs and number of streams. The outcome was somewhat expected as listening to whole playlists is relatively common among users.

6. Suggestions for next steps in analyzing data

There are some more aspects that can be examined using the dataset. First, the remaining hypotheses can be tested and more can be formulated. The impact of audio features on the number of streams can be examined. The correlation between key and valence of song can also be determined. Going beyond the dataset, the change of audio feature values throughout can be analyzed in more detailed way by adding more songs from different years.

7. Quality of dataset

The overall quality of dataset is considered to be high. There occurred some null values as well as inconsistent data types within one column. There was also mistake regarding the release year of the song. No duplicates occurred. The dataset has high usability.