CSCE 5215 - Machine Learning

Project Title: Most Streamed Spotify Songs 2023

Github Link: https://github.com/amireddy51/Most-Streamed-Spotify-Songs-2023

Team Members:

Venkateswarlu Amireddy - 11677056

Abinay Goud Karnam – 11665110

Padmaja Soma – 11626745

Yeshwanth Govindu – 11703977

Arun Kumar Gangireddy – 11656859

Abstract:

As of 2023, the Spotify charts are dominated by a kaleidoscope of musical genres, with the most played songs reflecting the platform's dynamic and ever-changing terrain. From pop choruses to hip-hop beats and indie jewels, the range of these tunes reflects the broad taste of Spotify's global audience. This data emphasizes the platform's function as a melting pot for musical innovation and cross-cultural influences.

The emergence of independent and foreign musicians is one of the key themes in the most played Spotify songs of 2023, showing the rising democratization of the music business. Collaborations involving performers from various genres and backgrounds, demonstrating the dissolving of conventional genre boundaries. Furthermore, the influence of viral challenges and social media on music popularity is clear.

Introduction:

In a time when music crosses boundaries and characterizes cultural events, Spotify is a leading indicator of the tastes of people throughout the world. The streaming platform remains the beating heart of the music industry, reflecting the pulse of what appeals to listeners all across the world, even as we make our way through the aural landscape of 2023.

The Most Streamed Spotify Songs of 2023" is a monument to the power of catchy melodies, meaningful words, and powerful rhythms in this ever-evolving musical journey. The fusion of genres and musicians reflects the varied fabric of modern musical expression, ranging from unexpected sleeper hits.

With a selection of songs that have dominated playlists, sound tracked events, and indelibly etched themselves into the collective memory of a global audience, this anthology not only honours the exceptional artistry of performers but also reveals consumer preferences.

Come explore with us the sounds, rhythms, and melodies that have reverberated throughout media, breaking over barriers to become the year's biggest hits and shaping 2023's musical landscape.

When it comes to music in the digital era, Spotify is the virtual stage where tunes from all over the world join together to create a musical mosaic that captures the spirit of the times. Music has become the soulful soundtrack to our lives. We discover an engrossing story of musical trends and cultural influences as we delve into the data-driven world of Spotify's most streamed songs of 2023.

The first step in discovering the songs that have been streamed on Spotify the most this year is to understand the complexities of data analytics. Spotify's analytic abilities, which are carefully crafted to record play counts, user interactions, and listener engagement, allow the site to compile an annual list of songs that have caught people's attention

The most streamed song playlist is a canvas decorated with a wide range of musical styles and tonalities. From radio-friendly pop hits that reach the top of charts to indie tracks that become cult favourites, the fusion of genres reflects the many interests and inclinations of a multicultural audience.

In addition to being melodies and beats, these songs are like time capsules that hold onto happy, resilient, activist, and introspective moments. They are voices that resound beyond the boundaries of a melody, amplifying voices that resonate with societal trends, and embodying the cultural zeitgeist.

Both creative genius and commercial success can be found in this carefully chosen collection. Every song is an artistic pursuit that has made a lasting impression on the musical landscape of 2023, from the heartfelt lyrics to the avant-garde soundscapes created by creative producers.

Let's explore the cultural legacies, inspirations, and backstories of the year's most streamed songs as we go on our journey through them, in addition to enjoying their success. Every song tells a story while evoking an emotional response that goes beyond data and unites us via the common language of music.

In addition to laying out the framework for discussing the Most Streamed Spotify Songs of 2023, this in-depth introduction explores the technical, cultural, and emotional aspects that characterize these songs, enticing readers to take a closer look at their relevance in the year's musical landscape.

Motivation:

Trend Analysis: Record labels, musicians, promoters, streaming platforms, and other businesses can all benefit greatly from having a solid understanding of the prevailing tastes and trends in music. Through this study, listener behaviour patterns, emerging genres, and factors that influence a song's appeal may be found.

Information for Music Labels and Artists: Artists and labels can gain important insights by being aware of the songs that receive the highest streams. By observing what appeals to consumers, they might gain insights that could inform their future marketing and music production techniques.

Displaying the most listened songs on Spotify or other streaming platforms can be a feature that draws users in. As a result, there may be an increase in user involvement and time spent on the site as listeners are encouraged to check out new music or revisit well-known songs.

Reporting on the songs that receive the most streams may be of interest to bloggers, influencers, and media outlets. This project can be useful for the production and analysis of material in the music business.

By producing visually appealing visual representations of the data, like charts or infographics, the information can be made more widely available. This can help to clearly and understandably communicate complex information concerning streaming music patterns.

Significance:

Determining the year's most streamed songs offers an overview of the time period's musical tastes and cultural climate. It displays the songs that are most popular with listeners and influence conversations about entertainment, music, and societal trends.

It's critical for the music industry to know which songs are most streamed. It provides insightful information on customer behavior, popular genres, and the efficacy of marketing tactics. This data helps record labels, musicians, and streaming services make well-informed choices about playlist curation, artist signings, promotion, and content production.

The career of an artist can be greatly impacted by ranking among the most streamed songs on a site as big as Spotify. It may result in more exposure and acknowledgment as well as more chances for performances, partnerships, endorsements, and overall success in the music business.

By promoting well-known songs and possibly influencing users' listening preferences, Spotify and other services can enhance user engagement by displaying the most streamed songs. User retention on the platform may rise as a result of this interaction.

The project makes available a dataset for a variety of analyses, such as finding trends, correlations, and patterns in the way that people listen to music. This data can be used by scholars, analysts, and music lovers to gain a better understanding of listener preferences and behaviour.

The most streamed songs can be the subject of articles, videos, and social media material created by media outlets, influencers, and content creators using this information. This advances the wider cultural dialogue on entertainment and music.

Objective:

The objective for analyzing the most streamed Spotify songs of 2023 is to gain insights into the contemporary musical landscape, identifying overarching trends, popular genres, and emerging artists. This analysis aims to provide a comprehensive understanding of the factors influencing music consumption on the Spotify platform, including the impact of cross-cultural collaborations, viral challenges, and the role of independent and international artists.

Additionally, the objective is to examine the patterns of user engagement and content discovery on Spotify, exploring how social media and user-generated content contribute to the popularity of certain songs. By delving into the most streamed songs, the goal is to uncover the preferences of a diverse global audience and to highlight the platform's role in shaping and reflecting cultural and musical trends.

Features:

Analysing the top Spotify tracks of 2023 entails taking into account a variety of factors that contribute to their popularity. Consider the following traits and factors:

Collaborations, Influence, Geographical Impact Artists, both independent and international Metrics Streaming, Chart Performance, UGC stands for User-Generated Content. Themes and Content in Lyrics Marketing and advertising

By extensively evaluating these qualities, we may acquire a comprehensive picture of the elements influencing the performance of the most streamed Spotify songs in 2023.

Dataset :

Related Work:

I don't have particular information regarding a "Most Streamed Spotify Songs 2023" dataset as of my last knowledge update in January 2023, because my training data covers information up to that point. The availability of such a dataset would be contingent on Spotify's releases or third-party compilations of data on the most-streamed songs.

Check official Spotify announcements, music industry sources, or popular music rankings for the most recent and relevant information about the most-streamed Spotify tracks in 2023. Furthermore, numerous data analytics companies and websites may create and publish lists of the most-streamed songs based on streaming data. Keep in mind that certain datasets may be subject to copyright or proprietary limitations.

The method's purpose and functionality are described in a docstring within triple quotes. Parameters are listed in the parentheses, and the method's return type and value are defined in the return statement. The actual implementation of the method follows, consisting of the code block inside the function. The method can then be invoked elsewhere in the code by its name, with input arguments passed as needed.

To show information about input characteristics in a dataset, provided a DataFrame called df1 exists. It loops over the DataFrame's columns, ignoring the "streams" feature, and outputs the serial number, feature name, and data type. The code closes by noting that the dataset's output feature is named "streams" and has a data type of int64. For the code to work properly, df1 must be replaced with the real DataFrame holding the dataset. We got columns description as

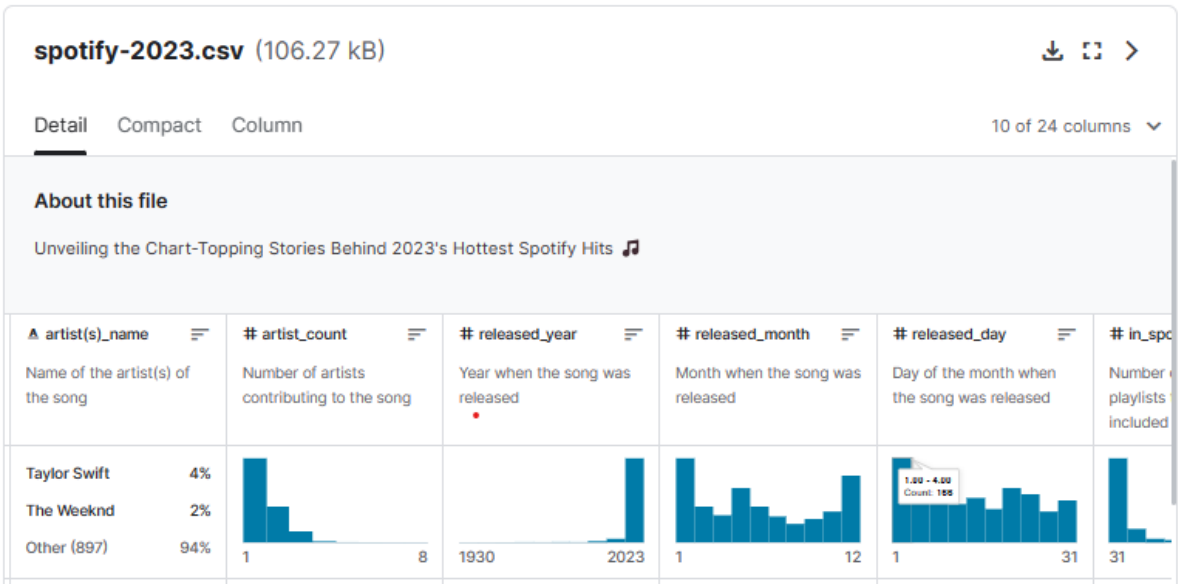
track_name: Name of the song

artist(s)_name: Name of the artist(s) of the song

artist_count: Number of artists contributing to the song

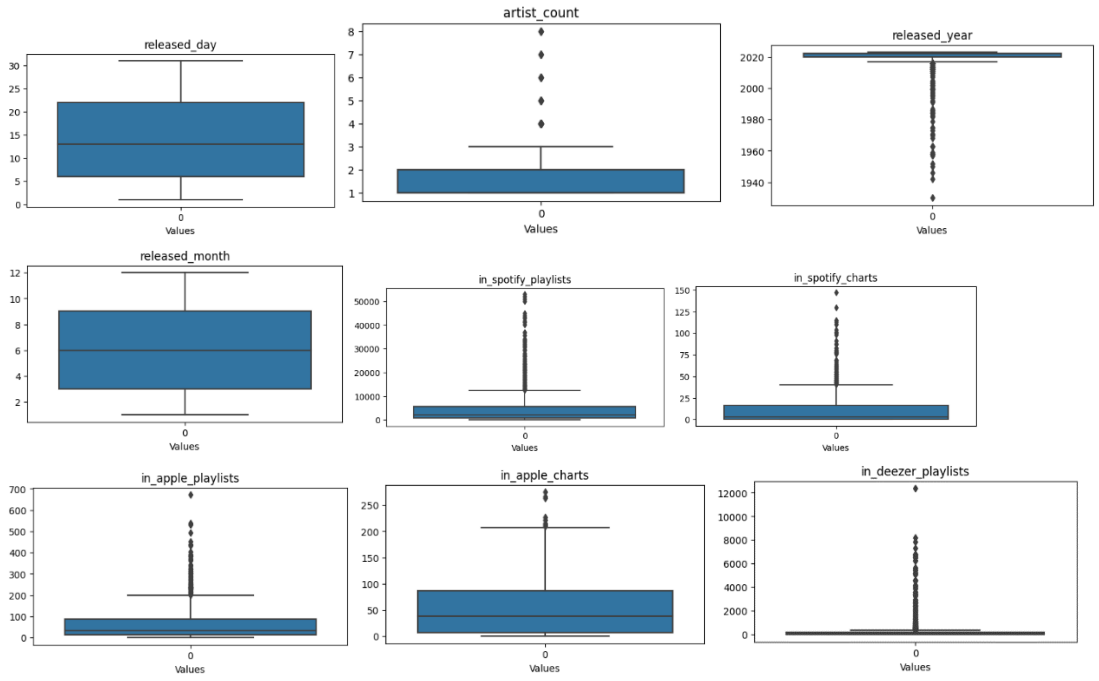
released_year: Year when the song was released
released_month: Month when the song was released
released_day: Day of the month when the song was released
in_spotify_playlists: Number of Spotify playlists the song is included in
in_spotify_charts: Presence and rank of the song on Spotify charts
streams: Total number of streams on Spotify
in_apple_playlists: Number of Apple Music playlists the song is included in
in_apple_charts: Presence and rank of the song on Apple Music charts
in_deezer_playlists: Number of Deezer playlists the song is included in
in_deezer_charts: Presence and rank of the song on Deezer charts
in_shazam_charts: Presence and rank of the song on Shazam charts
bpm: Beats per minute, a measure of song tempo
key: Key of the song
mode: Mode of the song (major or minor)
danceability_%: Percentage indicating how suitable the song is for dancing
valence_%: Positivity of the song's musical content
energy_%: Perceived energy level of the song
acousticness_%: Amount of acoustic sound in the song
instrumentalness_%: Amount of instrumental content in the song
liveness_%: Presence of live performance elements
speechiness_%: Amount of spoken words in the song

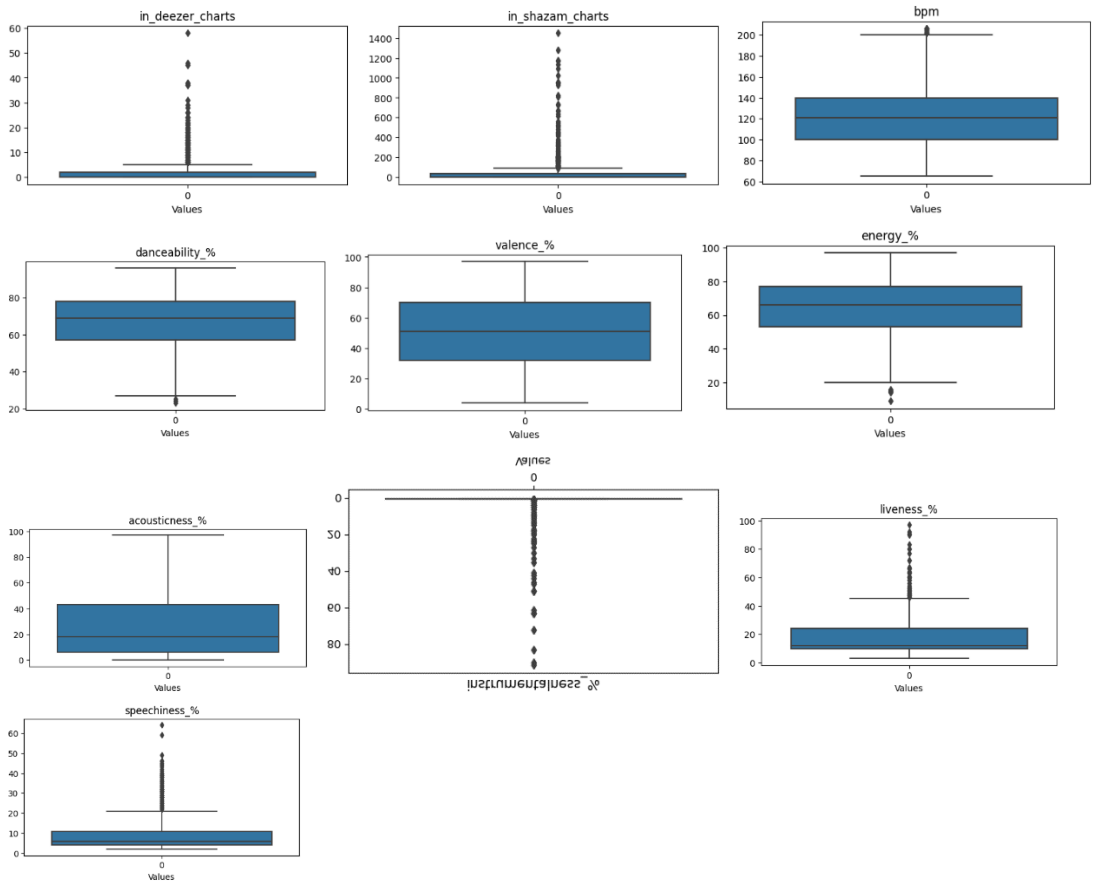
Dataset link: https://www.kaggle.com/datasets/nelgiriyewithana/top-spotify-songs-2023



as of now, in our datasets we have changed some columns into integers which have strings type of values and plotted the boxplots for numerical features from dataframe

Detailed Design of Methods:

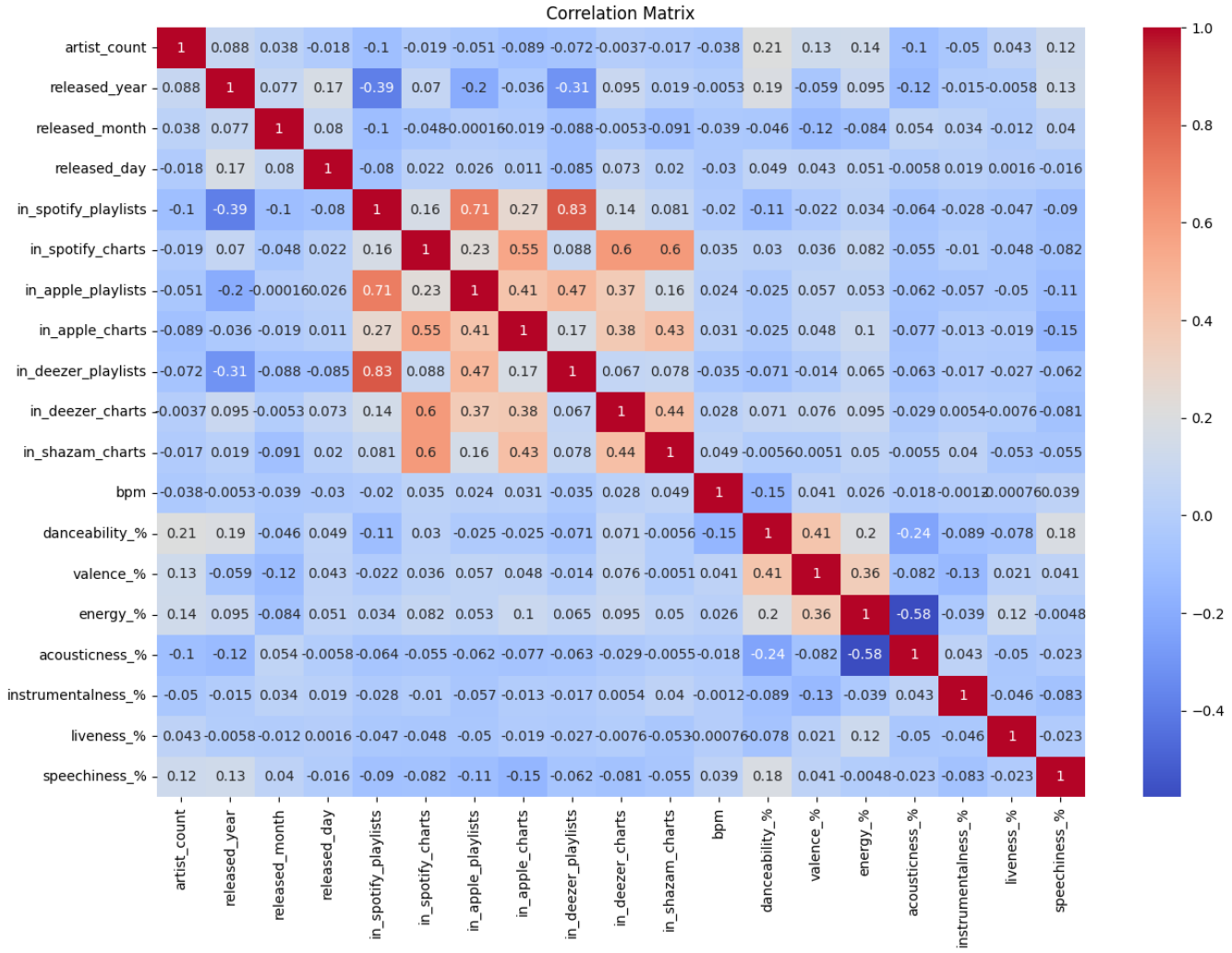




- It looks like most of the songs were create by 1 artist or a group of 2 artists there are very less songs which were created by group more than 2
- It looks like most of the songs in this data are from 2000 there are very less songs in this data which are before 2000
- It looks like most of the songs released in this data are released in between in january-febrauary or in between November-december
- It looks like most of the songs in spotify are there in between 1-10000 playlist after that there are very less songs which we can see where are there in more than 10000 playlists
- The playlist data in apple music is less compared to spotify playlists data and sezzer playlist data
- It looks like all the playlist data are following powerlaw distribution
- It looks like most of the songs in this data are using less than 10 words in their songs
- energy data is very close to fallowing normal distribution

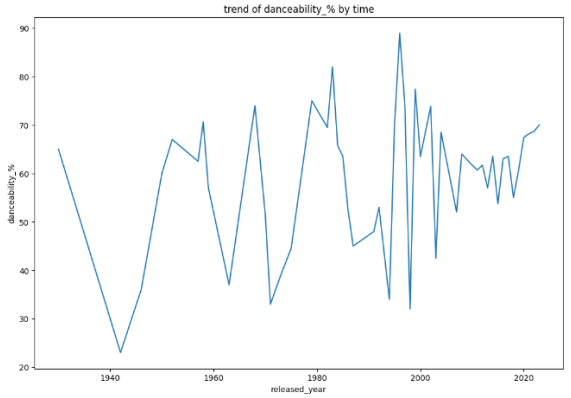
Analysis:

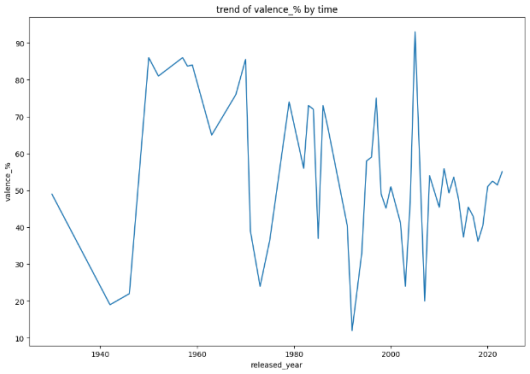
To get the correlation we have taken a heatmap and below is the image

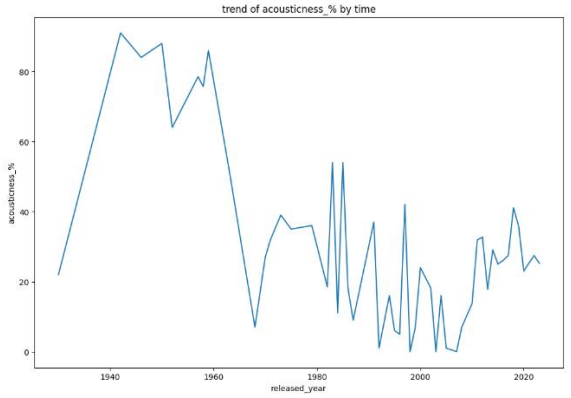


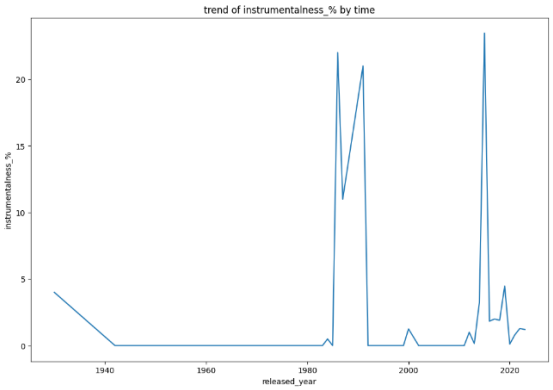
- All the data related to playlists like spotify playlists ,Apple playlists ,Dezzer playlists are positively corealted which tells that a song which is popular among playlists in one platform is popular in other platform playlists also
- All the data related to ranking like spotify charts,Apple charts ,Dezzer charts ,shazam charts are positively coralated which tells that a song popular in platform is most of the times popular in the other platforms also
- We can see a positive corelation between danceablity,valency,energy also which tells us a song with good valency may have high energy and dancebility value and vice versa
- accousticness and energy are highly negative correalted Which implies songs eith high accousticness have less energy levels in them
- And It looks like accousticness is negatively correalted with many other features in the data
- All the types of playlists data is negatively correlated with released year

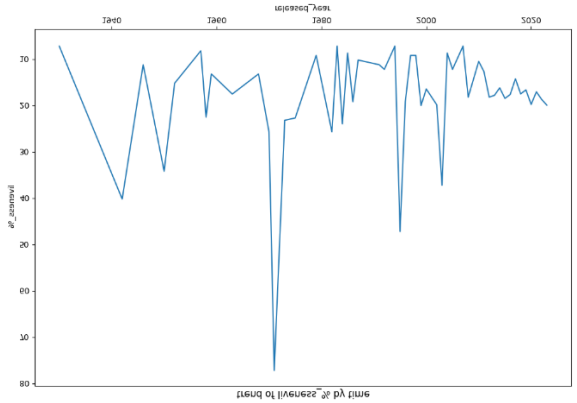
And later for the latest songs as per the time we have plotted the trend line plot for the song release year with the each feature.

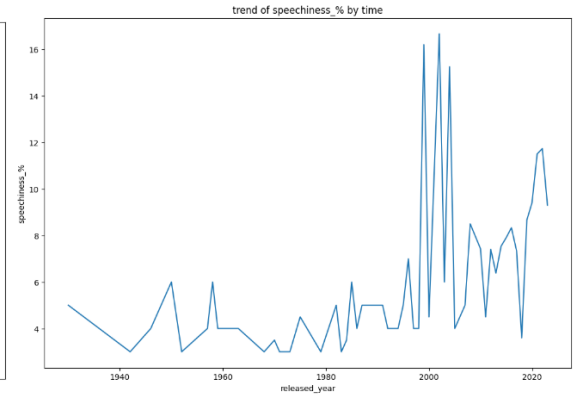


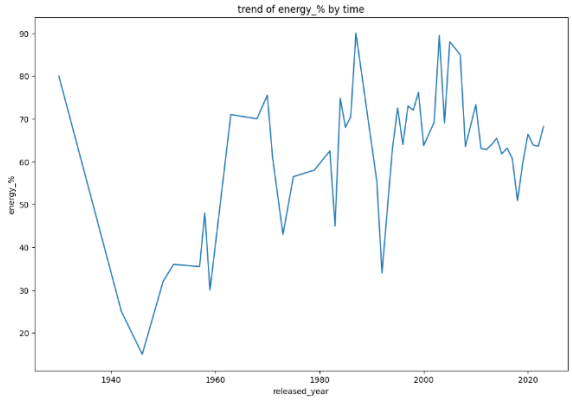












- As we can see from from the speechness graph as the time proceeded speechness increaes that is number of words in lyrics increased
- Accousticness in songs was popular during 1940 to 1960 after that accousticness in the songs decreased
- energy of the songs was low during 1940 but from 1960 energy is songs also incresed
- Intial upto 1980 Instrumental content in the songs was very less it took a boom in 1980 and again falled in 200 and took boom in 2020 for a small period of time
- Live performance element was very high during 1980 but before and after it it is decreased
- During 1980-2000 most of the song produced are suitable for dancing after that it decresed slightly before that there were many songs which were suitable for dancing but during this period many songs produced which can be used for dancing

➔ Encoded the column from string to numerical values as 1 as major and 0 as minor
➔ And taken the unique values by track name and artist names.

unique values in track_name: 941
unique values in artist(s)_name: 644
removing artist name and playlist because both columns unique value are very high which in a regression task is not useful

Implementation:

Model Training:

And later we have spitted the dataset into train data and test data. And filled the missing values by using the Simple Imputer, KNN Imputer.

One-hot encoding is a set of bits where the only permissible value combinations are those with a single high (1) bit and all others low (0). A comparable method in which all bits are '1' but one '0' is known as one-cold.

Model Selection:

For this dataset, we are using a Supervised machine learning approach where we use machine learning models like XG Boost, Linear Regression, Random Forest and with that we did a model evaluation of RMSE, RAE, R squared.

In this RMSE shows the model improves with decreasing RMSE, In RAE lower RAE denoted the higher model performance and In R-squared error a better match can be indicated by a greater R2.

Preliminary Results:

As per our models now the results are:

In Linear Regression, got the accuracy of 79%

R-squared (R^2) Score before hyper parameter tunning for test: 0.79
mean absoultue error before hyperparameter tunning: 139680254.97902098
mean squared error before hyperparameter tunning: 5.025688296810742e+16
root mean square error before hyperparameter tunning: 224180469.64021516
R-squared (R^2) Score for train before hyperparameter tunning: 0.73

In XG Boost, got the accuracy of 83%

mean absoultue error before hyperparameter tunning: 139680254.97902098
mean squared error before hyperparameter tunning: 5.025688296810742e+16
root mean square error before hyperparameter tunning: 224180469.64021516
R-squared (R^2) Score for test before hyperparameter tunning : 0.83
R-squared (R^2) Score for train before hyperparameter tunning: 1.00

In Random Forest (Bagging Model), got accuracy of 86%

Mean Squared Error: 5.025688296810742e+16
R-squared (R^2) Score before hyper parameter tunning for test: 0.86
mean absoultue error before hyperparameter tunning: 139680254.97902098
mean squared error before hyperparameter tunning: 5.025688296810742e+16
root mean square error before hyperparameter tunning: 224180469.64021516
R-squared (R^2) Score for train before hyperparameter tunning: 0.97

And later we did a hyper parameter tuning, because the difficulty of selecting an ideal collection of hyperparameters for a learning algorithm is known as hyperparameter optimization or tuning. A hyperparameter is a parameter whose value is utilized to regulate the learning process.

For XG Boost, we have seen a change of the accuracy learning rate to 85% and Linear Regression gives us a false intercept because of not improve in its testing accuracy. The results are as below:

XG Boost:

R-squared (R^2) Score before hyper parameter tunning for test: 0.85

mean absoultue error before hyperparameter tunning: 139680254.97902098
mean squared error before hyperparameter tunning: 5.025688296810742e+16
root mean square error before hyperparameter tunning: 224180469.64021516
R-squared (R^2) Score for train before hyperparameter tunning: 0.93

Best Hyperparameters:
1.learning rate: 0.1
2.max depth: 3
3.max_leaves: 5
4.n_estimators: 100

Linear Regression:

R-squared (R^2) Score before hyper parameter tunning for test: 0.79
mean absoultue error before hyperparameter tunning: 139680254.97902098
mean squared error before hyperparameter tunning: 5.025688296810742e+16
root mean square error before hyperparameter tunning: 224180469.64021516
R-squared (R^2) Score for train before hyperparameter tunning: 0.73

Best Hyperparameter:
fit_intercept: False

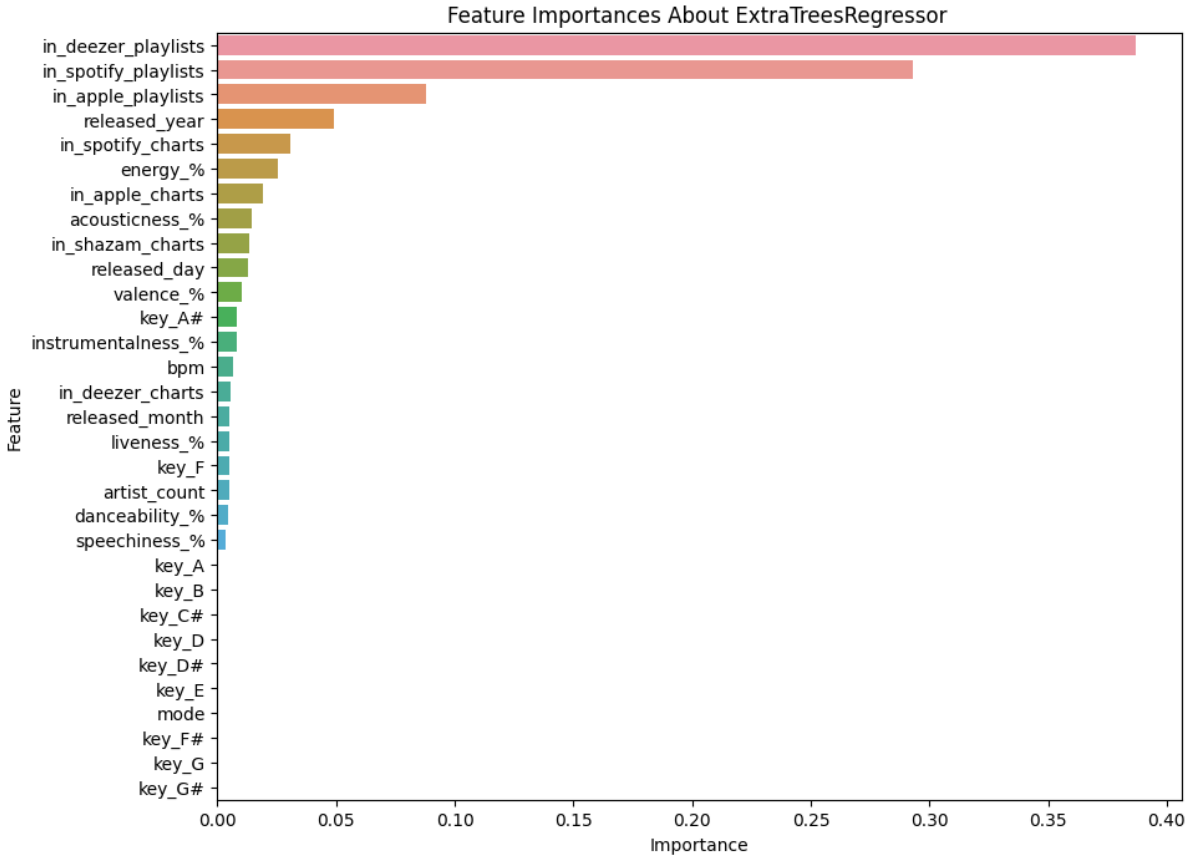
For Random Forest (Bagging Model):

R-squared (R^2) Score before hyper parameter tunning for test: 0.87
mean absoultue error before hyperparameter tunning: 139680254.97902098
mean squared error before hyperparameter tunning: 5.025688296810742e+16
root mean square error before hyperparameter tunning: 224180469.64021516
R-squared (R^2) Score for train before hyperparameter tunning: 0.97

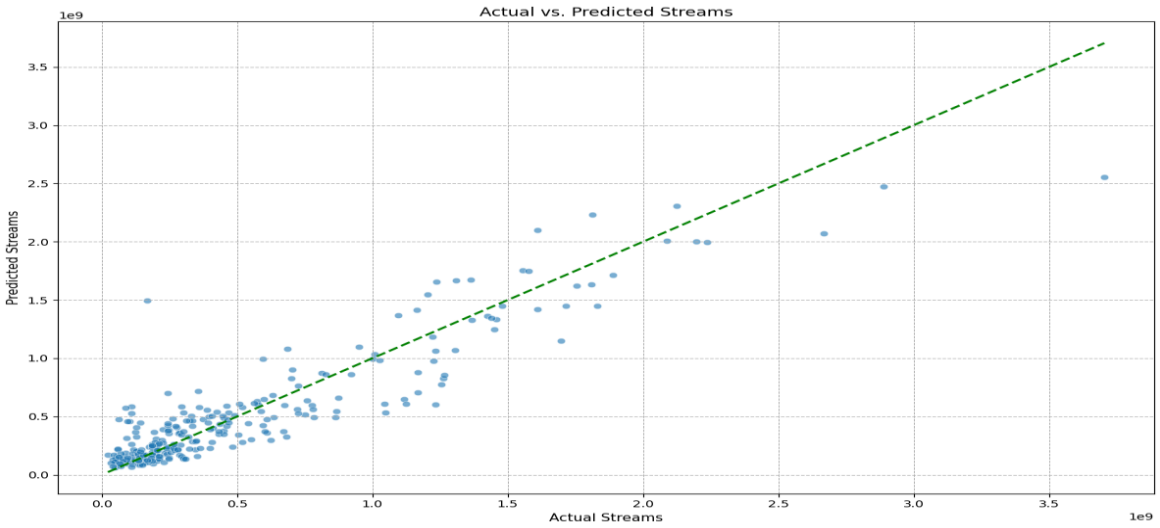
Best Hyperparameters:
1.n_estimators: 70
2.max depth: 12
3.max_sample_split: 2
4.min_sample_leaf: 4
5.max_features: auto

In the overall 3 models, we have got the best accuracy for Random Forest, that is 87%.

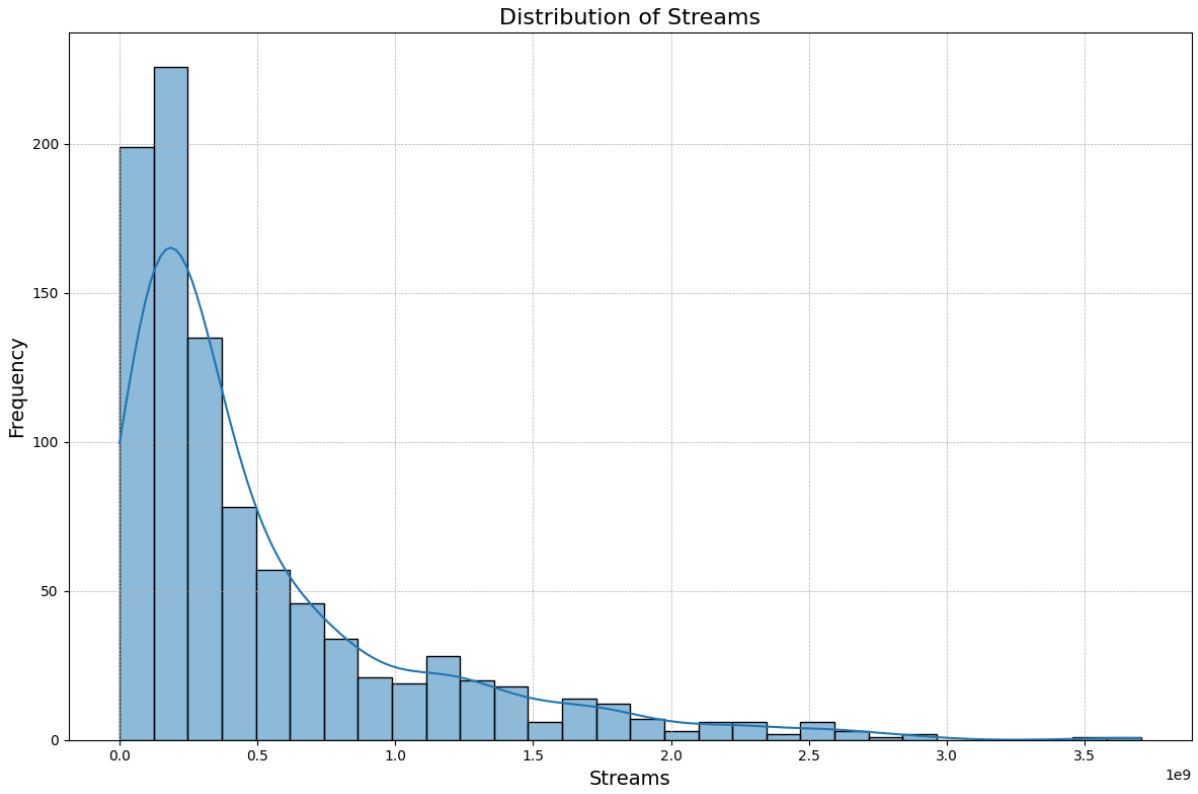
- ➔ GridSearchCV yielded an XGBoost model for extracting and visualizing feature significance. The feature importances are saved in a DataFrame and sorted in decreasing order, providing a clear depiction of each feature's relative importance. The resultant bar plot, built with matplotlib and seaborn, offers a visual summary of how different characteristics contribute to the model's predictions. The y-axis labels the traits, while the x-axis shows their relative importance. The plot is headed "Feature Importances About ExtraTreesRegressor," suggesting its importance in understanding the influence of features on the performance of the XGBoost model.



➔ A scatter plot comparing real streams (Y_test) to predicted streams (Y_pred1) after grid search with cross-validation. The figure size is set to 12 by 8 inches, and a scatter plot with transparency (alpha=0.6) is constructed for better viewing of data points. A green dashed line is drawn to indicate the ideal condition in which the actual and anticipated values are equal. The x-axis shows actual streams, the y-axis represents expected streams, and the graphic is named "Actual vs. Predicted Streams." Grid lines are used for clarity, and the layout is tweaked to look nicer. This graphic enables for an assessment of how well the model predictions correlate with the actual values, offering insights into the model's performance.



To show the distribution of the 'streams' variable in the DataFrame, a histogram with a kernel density estimate plot is used. The histogram and bin edges are calculated using the np.histogram function, and the histogram plot with the provided number of bins and the overlay of kernel density estimate is created using the sns.histplot function. The variable 'streams' is shown on the x-axis, the frequency of occurrences is represented on the y-axis, and the plot is named "Distribution of Streams." Before presenting the plot, grid lines are added for greater reading, and the arrangement is changed for improved look. This graphic sheds light on the spread and form of the dataset streams variable distribution.



Result

We have used a machine learning approach and used 3 models of approach of XG Boost, Linear Regression, Random Forest (Bagging Model). Got the accuracy of 83%, 79%, 86% respectively. After hyper parameter tuning 85%, 79%, 87% respectively.

Project Management:

Implementation Status Report

Work Completed:

Description:

Feature Selection, Model Selection, Model training and hyper parameter tuning was done for models XG Boost, Linear Regression, Random Forest (Bagging Model), Model Evaluation have been completed till now.

Responsibilities:
Venkateswarlu Amireddy – About the Dataset, Exploratory Data analysis

Abinay Goud Karnam – Detailed Design of methods, Analysis and gave contribution in making the report.

Padmaja Soma – model selection and model training

Yeshwanth Govindu – implemented two models and gave contribution in making the presentation

Arun Kumar Gangireddy – Implemented one model and validation of three models, also gave contribution in making the report

Contributions:
Venkateswarlu Amireddy – 20% - Exploratory Data analysis and helped in making report

Abinay Goud Karnam – 20% - Detailed Design of the of methods, Analysis, helped in the EDA part and plotted different graphs and gave contribution in making the report.

Padmaja Soma – 20% - helped in the model selection and the model training models

Yeshwanth Govindu – 20% - Helped in the model selections and explained.

Arun Kumar Gangireddy – 20% - Helped in contribution of getting models trained and tested

Responsibilities:

Issues and concerns:

Linear regression is susceptible to data outliers. Outliers can have a disproportionate impact on the model's coefficients, resulting in erroneous predictions.

Depending on the complexity of the model and the number of characteristics, linear regression may either overfit or underfit the data. If the model is too simplistic, it may fail to detect complicated patterns in the data.

References:

https://kworb.net/

https://newsroom.spotify.com/

https://twitter.com/Spotify

Conclusion

A distinct trend develops, with a preference for single performers or duos over bigger groups. The collection has a modern bent, with music largely published after 2000. Release frequency peaks are notable in January-February and November-December, indicating possible strategic timing in the music business. The popularity distribution shows a steep drop in playlist inclusion beyond 10,000, with Spotify playlists greatly outnumbering those on Apple Music and Shazam. Surprisingly, the energy levels across songs closely coincide with a normal distribution, implying a balanced energy distribution across the sample. This in-depth examination offers insight on the complex qualities and distribution tendencies of 2023's most streamed Spotify tracks.

This project is concerned with determining whether or not a song will become popular.

Artist Insight and Decision-Making: By developing an accurate forecasting model, artists may learn a lot about the future popularity of their music. Because of this information, they are better prepared to make judgments about their artistic activities.

Allocation of Funding and Resources: Using this approach, artists may decide whether a song has a decent possibility of becoming successful.

Radio Channel Programming: Radio stations can use this approach to estimate the potential popularity of songs before deciding whether or not to include them in their playlists. This ensures that the content they provide for their audience is relevant to their interests and inclinations.