Spotify Tracks Analysis

**INTRODUCTION**

Spotify has over 248 million monthly active users across the globe for digital podcasts and music streaming services. The company's subscription service called 'Spotify Premium' already has an impressive rate of rise in its customer base of +31% (year on year growth). While its important features indicate that an average customer spends 25 hours a month on the service, it is equally fascinating to dig in and learn from the data behind the scenes. This 'king of streaming music' is well known for its custom music recommendations for its users and our following analysis examines the main determinants that affect Spotify's success. The analyses are mostly aimed at helping companies and music dealers operating within the field of digital streaming services.

**Problem Statements**

There are two main objectives for this analysis:

* Identify Spotify users' general artistic affinity patterns
* Determine the key factors of Spotify track success

**Techniques Used**

The dataset includes details about the artist, the genre, features and popularity of Spotify's songs. To achieve our goals, the cleaned data will be evaluated using EDA techniques and data mining techniques (for example regression, word tokenization and clustering). The variable importance for track popularity can be defined by regression, such as Random Forest / Multiple Linear Regression. Tokenization and NLP approaches will help to determine if a collection of terms (in the titles) for more common tracks can be found.

**Key Consumers**

The analyses are intended specifically to support businesses and music providers working in the area of digital streaming services. Digital music publishers may help define consumer tendencies and enhance their music offerings. The study will also allow artists to better identify the target audiences (as the analysis is split by music genre).

**Source Data**

There are 32833 observations and 23 variables of the source results. The dataset includes 15 missing values and the initial data set did not have these values. The vector "track id" is a single song id, while 4477 duplicate values are displayed, as one song on the Spotify dataset can be linked to multiple genres. The variables involved for consideration are summarized below. 'track popularity' is one of the major interest variables, and the variable is summarized statistically. In the EDA method, variables are not deleted, so the definition of all these variables is given below:

|  |  |  |
| --- | --- | --- |
| Variable | Class | Description |
| track\_id | char | Song Unique ID |
| track\_name | char | Song Name |
| track\_artist | char | Song Artist |
| track\_popularity | double | Song Popularity (0-100) where higher is better |
| track\_album\_id | char | Album unique ID |
| track\_album\_name | char | Song album name |
| track\_album\_release\_date | Factor | Date when album released |
| playlist\_name | char | Name of playlist |
| playlist\_genre | char | Playlist genre |
| playlist\_subgenre | char | Playlist subgenre |
| danceability | num | Danceability describes how suitable a track is for dancing. A value of 0.0 is least danceable and 1.0 is most danceable. |
| energy | num | Energy is a measure from 0.0 to 1.0 and represents a perceptual measure of intensity and activity. |
| key | int | The estimated overall key of the track. |
| loudness | num | The overall loudness of a track in decibels (dB). Loudness values are averaged across the entire track and are useful for comparing relative loudness of tracks. |
| mode | int | Mode indicates the modality (major or minor) of a track, the type of scale from which its melodic content is derived. Major is represented by 1 and minor is 0. |
| speechiness | num | Speechiness detects the presence of spoken words in a track. The more exclusively speech-like the recording (e.g. talk show, audio book, poetry), the closer to 1.0 the attribute value. Values above 0.66 describe tracks that are probably made entirely of spoken words. Values between 0.33 and 0.66 describe tracks that may contain both music and speech, either in sections or layered, including such cases as rap music. Values below 0.33 most likely represent music and other non-speech-like tracks. |
| acousticness | num | A confidence measure from 0.0 to 1.0 of whether the track is acoustic. 1.0 represents high confidence the track is acoustic. |
| instrumentalness | num | Predicts whether a track contains no vocals. “Ooh” and “aah” sounds are treated as instrumental in this context. Rap or spoken word tracks are clearly “vocal”. The closer the instrumentalness value is to 1.0, the greater likelihood the track contains no vocal content. Values above 0.5 are intended to represent instrumental tracks, but confidence is higher as the value approaches 1.0. |
| liveness | num | Detects the presence of an audience in the recording. Higher liveness values represent an increased probability that the track was performed live. A value above 0.8 provides strong likelihood that the track is live. |
| valence | num | A measure from 0.0 to 1.0 describing the musical positiveness conveyed by a track. Tracks with high valence sound more positive (e.g. happy, cheerful, euphoric), while tracks with low valence sound more negative (e.g. sad, depressed, angry). |
| tempo | num | The overall estimated tempo of a track in beats per minute (BPM). In musical terminology, tempo is the speed or pace of a given piece and derives directly from the average beat duration. |
| duration\_ms | num | Duration of song in milliseconds |

**DATA ANALYSIS**

1. Average genre ranking and number of spotify tracks

Key Insights:

* Although the "pop" genre has less songs on spotify, its popularity score is the best among consumers.
* 'edm' and 'rap' have more songs on spotify, hence emerging artists in these genres compete in terms of their track exposure more rigorously.

A picture containing chart

Description automatically generatedChart

Description automatically generated

1. Correlation Matrix

Key Insights:

* With a higher song duration and high instrumentality, the track success declines marginally.
* The danceability of the track is favorably connected to a track's Valence (positivity).

Chart, bubble chart

Description automatically generated

1. Podcast vs Music tracks

Key Insights:

* Podcasts are less common on average than spotify music tracks, although the number of podcasts is much smaller than the number of music songs. For the reason there is also a clear chance to extend the customer base and the options for Spotify podcasts.

![Text

Description automatically generated]()

1. Popularity changes over the age of album release

Key Insights:

* Early release rock music is more common with users than the latest tracks
* "R&B" genre has become increasingly popular with spotify users, with the recently released 2010 songs being higher than the 2000, the 90s and 80s albums.

1. Popular Genre Identification

Key Insights:

* In terms of their average popularity the recent tracks 'pop' and 'latin' have a higher ranking.
* The latin music published in '2010s' is much more recognizable and popular among spotify users compared to 2000's.
* 2000

![Chart, bar chart

Description automatically generated]()

* 2010

![Chart, bar chart

Description automatically generated]()

1. Commonly appearing song titles by popularity

Key Insights:

* There is a disparity in mainstream music in the most frequent titles vs. less popular ones. Titles having words 'dance,' 'hits,' 'hop,' are more common in highly popular songs than in less popular tracks.

**SUMMARY OF ANALYSIS**

This research focuses primarily on providing key patterns and insights in the field of digital streaming platforms for music artists and distribution. For different types of music and various releases, general patterns in the success of tracks were established. The main elements of similarities were contrasted with less common tracks in terms of "title words." Packages like Plotly have been used in an immersive way to reflect main patterns. Text mining techniques have been used to create track title word clouds .The analyses are mostly intended to assist corporations and music distributors operating within the DSD (digital streaming domain). Identifying consumer patterns will boost content supply to digital music distributors. The research will also allow artists to develop their perception of customers (as the analysis is split by music genre).

**LIMITATIONS**

* A static data collection is used to analyze the results. This could be augmented by the dynamic scrapping of data through a web API
* If available, user feedback can be used to explain the main aspects that enhance the user's association with music/podcast tracks.
* The track titles are not consistent.