

**M9.9 Final Project-Spotify**

**project m9.9**

December 15, 2021

taran kumar

BIA-610 APPLIED ANALYTICS

Stevens Institute of Technology – Summer 2021

Table of Contents

[1 Spotify exploratory data analysis 2](#_Toc122038787)

[1.1 INTRODUCTION 2](#_Toc122038788)

[1.1.1 Problem Statement 2](#_Toc122038789)

[1.1.2 Solution Overview 3](#_Toc122038790)

[1.1.3 Insights 3](#_Toc122038791)

[1.2 Data Preparation & Data Cleaning 4](#_Toc122038792)

[1.2.1 Data Cleaning 5](#_Toc122038793)

[1.2.2 Data Preview 6](#_Toc122038794)

[1.3 Exploratory Data Analysis: 7](#_Toc122038795)

[1.4 Conclusion/Summary 17](#_Toc122038796)

Table of Figures

Figure1: Data Preview (Tableau)………………………………………………………….. 6

Figure2: Album Release Date Vs Track Popularity………………………………………...7

Figure3: Track Album Name Vs Energy…………………………………………………...8

Figure 4: Track Album Release Date Vs Valence & Duration of song…………………….9

Figure 5: Top 100 Songs Vs Liveness…………………………………………………….10

Figure 6: Track Name Vs Danceability…………………………………………………... 11

Figure 7: Track Artist Vs Popularity………………………………………………………12

Figure 8: Danceability Vs Loudness of Track…………………………………………….13

Figure 9: Danceability Vs Track Popularity………………………………………………14

Figure 10: Genre Vs Tempo……………………………………………………………….15

Figure 11: Top Tracks Vs Measure Values………………………………………………..16

# spotify exploratory data analysis

## INTRODUCTION

Spotify was founded in 2006 by Daniel Ek and Martin Lorentzon. It is one of the largest music streaming service providers and has more than 456 million monthly users active on their application. As of September 2022 report, they have 195 million people who have subscribed to it and pay monthly for it. The company follows a freemium business model. This model means, the company provides a free platform for the users to stream music for free with advertisements in between. Once the user gets used to it and like the platform, they can become a paid subscriber and listen to music without advertisement. Spotify has an algorithm that monitors the music that the individual user is streaming on daily basis and generates a recommendation to the individual according to their stream. This makes the user gets more obsessed with the platform which increases the number of subscribers. Spotify also applies machine learning programming to classify different music pieces, segregate similar tracks and generate delightful customer experiences. Spotify is doing good but there is a still long way ahead for the company. They have many competitors such as Apple iTunes. Spotify pays a large amount of its money to music labels which reduces its revenue and can give them a hard time

### Problem Statement

It is not easy to manage such a big company, it has huge programming and algorithms involved it as mentioned above. In my research, I am trying to answer a few questions about Spotify. The first thing is to determine what factors are impacting the popularity of the track and in this, we are trying to analyze the top 100 tracks on Spotify. Secondly, which type of songs are played the most, and in that which genres songs are most played frequently? Lastly, the aim is to analyze the attributes of the music which impact its popularity.

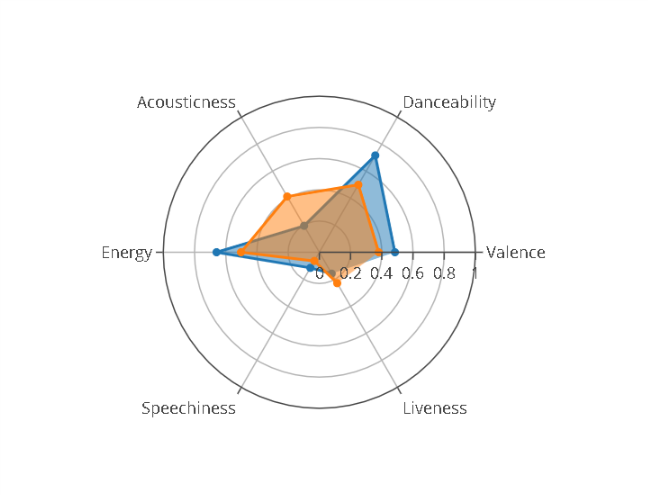
The primary goal of this project is to find answers to the questions that I have mentioned in the above paragraph with the help of an exploratory analysis of the Spotify songs.csv file. I believe before starting any kind of research it is important to understand the database and the reason why people like music in their daily life.

I will use Tableau in this project as a tool to answer the questions to our problem statement. We will try to figure out and look into different aspects of the dataset to see who the most popular artist is, and which songs are the most popular. This will also help us to know which artists are at the top.

### Solution Overview

Now, to answer all these issues mentioned in the problem statement, I have used the data inquiry (both multivariate and univariate), data cleaning process, and data analysis to eliminate and identify potential problems. To do this, I have used a mixture of python and excel. A lot of research has been done and dashboards along with photos are used in the analysis where appropriate to do this study.

### Insights

Several significant findings were made from the exploratory inquiry which are further addressed below in detail. Starting with the popular songs on Spotify, the fat Freestyle from the album 3rd coast classics received the most minutes of playtime. This means users played these songs most of the time(minutes), which shows how popular it is. Secondly, the most well-known musical genres in the playlist of the database are rock, EDM, and R&B. Other than this, we will see in the data visualization below that the frequency of music being played is substantially decreasing. Now, if we see the number of minutes, Maryin Garrix’s songs are listened to the most by the users compared to other performers.

## Data Preparation & Data Cleaning

The original dataset for Spotify was obtained from GitHub. There are 23 variables in the total of music listed on Spotify’s platform provided to us, along with 32,834 records on the different variables of music. The columns included in the database are the following….

Track Id: Track Id identifies each music track as the primary key and is unique.

Track Name: The track name shows the name of the track with a one-line sentence of the song.

Track Artist: This shows the name of the artist.

Track Popularity: This shows the popularity of the track which means how popular the track is between 0 to 100.

Track Album Id: This shows the unique Id of the album to which the track is connected.

Track Album Name: This shows the name of the track album.

Track Album Release Date: This represents when the track album was released.

Playlist Name: This shows what kind of music it is like the pop remix, rock, 70s and etc.

Playlist Id: This shows the unique Id associated with the playlist.

Playlist Genre: Genre is the category of artistic composition. This playlist genre shows different categories.

Playlist Subgenre: This shows a category within the category of the playlist genre.

Danceability: This shows the value between 0 to 1 which determines the suitability on which a person could dance to a part of a song or the whole song.

Energy: This also shows a value between 0 to 1 and it represents the power of the track.

Key: This shows the keynote of the track.

Loudness: This shows how loud the song is, this basically means the intensity.

Mode: Mode shows whether the song is major or minor.

Speechiness: This shows how many words are spoken in the track.

Acousticness: Acousticness is measured from 0 to 1 and shows whether the track is acoustic. Acoustic means the sense of hearing a song.

Instrumentalness: This shows non-vocal /no spoken words in the track Liveness: This shows the liveness/ performance of the music.

Valence: This shows sentiments in the track. How positive or negative it is.

Tempo: It shows the beats per minute of the track.

Duration\_ms: This shows the duration of the track.

### Data Cleaning

Data cleaning is a very important part of research and should be done before starting any project. It is the process of fixing or removing the data, which is incorrect, corrupted, incomplete, or incorrectly formatted in the database. In this research, the data cleaning process was carried out to make sure that the data is error-free before using it for analysis and visualization. Error-free means data is free of any missing or null entries. Microsoft Excel and Tableau were used to clean the dataset. The conditional formatting tool was used in excel to clean this dataset and it shows that there is no existence of missing or null values in the dataset. Moreover, the track name and track album name had the same names, so it was decided to delete one column so that data get easy to interpret and prevent it from generating unexpected outcomes. The track artist and track album name had a few low values also. Lastly, there were no duplicate values in the dataset which I believe is a good sign for analysis to do analysis.

### Data Preview

Graphical user interface, table

Description automatically generatedThe dataset is imported into tableau to do the analysis with the help of visualization. In this, each row contains the track, and each column has the information related to the track. To do the analysis, first we have imported the excel data into tableau and clicked on cleaned with data interpreter to make sure it is cleaned from the errors.

Figure1: Data Preview (Tableau)

## Exploratory Data Analysis:

**Track Album Release Date vs Track Popularity:**

Graphical user interface, text, application

Description automatically generated As mentioned above tableau will be used to do all the visualizations of this research. At first, I have taken into consideration the data of 20 years from 2000 to 2020 and filtered it based on their release date. We can see that there is not a great change in the graph from 2000 to 2013. After 2013, the graph started going on upscale and it has the highest popularity in the year 2019, followed by 2018. The 2019 year was very successful for the Spotify company as many albums were published in that year compared to other years. This also means that it was an important year for the company in view of its revenue for that year. The reason for the highest popularity in the year 2009 followed by 2008 for Spotify was that the company launched its Spotify service in 2008. The service offered listeners access to a vast library of recordings rather than asking them to pay for the download or for the CDs. However, it didn’t guarantee any change in revenue.

Figure2: Album Release Date Vs Track Popularity

**Track Album Name Vs Energy:**

In second one, I have started comparing the Track Album Name with its energy. We can see below that the Track Album name Greatest Hits has the biggest energy level on Spotify. The energy level for this album Greatest Hits is 96.65 and it is followed by the Ultimate Freestyle (Having an energy level of 30l19) and Appetite For Destruction album names (Having an energy level of 26.47). Here it is important to see that albums 2 and 3 are significantly different from album 1 (Greatest Hits). There is a huge difference in level between them. Secondly, we can see that only the greatest hits has the highest level of energy and all the rest have a small difference between their energy levels

Graphical user interface

Description automatically generated

Figure3: Track Album Name Vs Energy

**Track Album Release Date Vs Valence & Duration of song**

The reason for doing an analysis on this was to see what the valence of track albums is. I have covered it just under the track popularity so that we can easily compare them. The track album released in 2003 has the highest valance that is 0.6248, and an average duration of 246.296Ms. Moreover, the valance for every album is positive that shows songs introduced in the market are with positive sentiments which is a good thing.

Chart, line chart

Description automatically generatedFigure 4: Track Album Release Date Vs Valence & Duration of song

**Top 100 Songs Vs Liveness**

Here, I have designed a relationship between the track name and its liveness. I have limited the data to the top 100 tracks so that it is easy for the company to visualize. Liveness is basically the presence of an audience in a recording. In our case, we can see that poison has the highest liveness rating which is 6.47. This shows this track has been performed live a lot of time. On the second is The Box with 6.32 and the list goes on till 100 tracks. This information shows that which songs are performed live more so Spotify can use this data to make changes to their app accordingly

Chart, bar chart

Description automatically generated

Figure 5: Top 100 Songs Vs Liveness

**Track Name Vs Danceability**

The purpose of this research is to find out which songs have the most potential to dance to. This will also help to know their popularity as if more people dance to a song, it is going to get popularity. Based on different parameters such as rhythm, beat power, pace, and general regularity. We can conclude after looking at the below screenshot that dance monkey is the most appropriate tune for dancing. It has a rating of around 14 and its closest competitor is Breathe monkey, which has a danceability rating of 13.45. This research will help the company to make decisions that which type of songs are attracting more people to dance and then the company can promote them more or can charge extra for it if required.

Chart

Description automatically generated

Figure 6: Track Name Vs Danceablility

**Track Artist Vs Popularity**

Here I did another exploratory analysis of the database of Spotify. This is an association analysis that combines Track popularity and Track Artist. We can see in the screenshot below that Martin Garrix is the most famous artist as he has huge track popularity and is followed by The Chain Smoker. They both have the popularity of around 7k.

Chart

Description automatically generated with low confidence

Figure 7: Track Artist Vs Popularity

**Danceability Vs Loudness of Track**

Now as I have discussed above the track albums and track names are in detail along with their different parameters such as energy, liveness, popularity and etc. It is important to analyze the danceability and loudness of the tack also. This is because they both are connected to each other and they are an important element of the songs. If the song has high danceability, the loudness of the track will also be high which will make the audience go crazy and more addicted to that music. In this case, I have taken into account the years from 2000 to 2020. We can see that in 2016, the danceability and loudness is high.

Chart, scatter chart

Description automatically generated

Figure 8: Danceability Vs Loudness of Track

**Danceability Vs Tack popularity**

Here I have used clustering. Clustering is used to show similar numerical values. I have tried to compare danceability and track popularity using the clustering technique. The data below is divided into 2 clusters (1 & 2). Moreover, I have added parameters like energy, track name, and acousticness. This will give an overall observation of track popularity. If we see, Dreams- 2004 Remaster has the highest danceability, energy, and track popularity in cluster 2. In cluster 1, A horse with no name has the highest of these parameters.

Chart, scatter chart

Description automatically generated

Figure 9: Danceability Vs Track Popularity

**Genre Playlist Vs Tempo**

Chart, bar chart

Description automatically generatedHere I have compared the genre playlist with Tempo. We can see below that EDM (Electronic Dance Music) genre has the fastest pace. Its tempo is around 760,016 BMP. The music expressed here is in BMP (beats per minute). Moreover, tempo means the speed or pace of the piece of music. Tempo basically includes the average beat length of the composition of music.

Figure 10: Genre Vs Tempo

**Top Tracks Vs Measure Values**

Graphical user interface, chart

Description automatically generatedLastly, I have tried to do an overall analysis of the top 100 track titles. This includes the different information about the tracks such as their energy level, loudness, liveness, mode, tempo and speechness. Every track name has its own variation in the parameters and is presented in distinct way.

Figure 11: Top Tracks Vs Measure Values

## Conclusion/Summary

Data analysis is not an easy task and needs a lot of research and time. It is not possible to take into consideration all the parameters for the analysis in one study. So, this study was limited to exploratory data on Spotify music. I believe the research explained above is worthwhile as it provides new key insights to understand the Spotify music data in depth and more clearly which previously was not available like this. We can clearly see now that every artist has their own set of qualities in their songs which changes their popularity. Some people like loud music and some people like the music to which they can dance. Moreover, the data shows that EDM, Rap, Pop, R&B, rock, and Latin are the most popular playlist genre of all time. It also shows which artist has the most popularity and which tracks have the energy, danceability, and liveness in their songs.

This will eventually help the Spotify company to make decisions for their business. They can do more advertisements on the songs which are played more. Moreover, they can also do promotions according to the data that they have now received with the help of this analysis.

Lastly, I would say that this is not the end and there are many other parameters or ideas that can be taken into consideration while doing an in-depth analysis on this data. I hope to continue my research and would like to get more insights on it that can help the company and can improve my knowledge about it.

