**Data Set Title**

**Exploratory Analysis**

Lyn Detwiler, [ldetwiler@bellarmine.edu](mailto:ldetwiler@bellarmine.edu)

Garrett Ward, [gward2@bellarmine.edu](mailto:gward2@bellarmine.edu)

1. **INTRODUCTION**

This data set shows the most popular songs that have been streamed on Spotify. It includes the artist, release date, total number of streams, and how many playlists the song is apart of. We chose this data set because we both use Spotify and have a shared interest in music. The data set can be found here: <https://www.kaggle.com/datasets/abdulszz/spotify-most-streamed-songs>

1. **DATA SET DESCRIPTION**

This data set contains 953 samples with 25 columns with various data types. A complete listing in shown in **Table 1.**

**Table 1: Data Types and Missing Data**

|  |  |  |
| --- | --- | --- |
| *Variable Name* | *Data Type* | *Missing Data (%)* |
| in\_shazam\_charts | object | 5.25% |
| key | object | 9.97% |
| track\_name | object | 0.00% |
| artist(s)\_name | object | 0.00% |
| artist\_count | int64 | 0.00% |
| released\_year | int64 | 0.00% |
| released\_month | int64 | 0.00% |
| released\_day | int64 | 0.00% |
| in\_spotify\_playlists | int64 | 0.00% |
| in\_spotify\_charts | int64 | 0.00% |
| streams | object | 0.00% |
| in\_apple\_playlists | int64 | 0.00% |
| in\_apple\_charts | int64 | 0.00% |
| in\_deezer\_playlists | object | 0.00% |
| in\_deezer\_charts | int64 | 0.00% |
| bpm | int64 | 0.00% |
| mode | object | 0.00% |
| danceability\_% | int64 | 0.00% |
| valence\_% | int64 | 0.00% |
| energy\_% | int64 | 0.00% |
| acousticness\_% | int64 | 0.00% |
| instrumentalness\_% | int64 | 0.00% |
| liveness\_% | int64 | 0.00% |
| speechiness\_% | int64 | 0.00% |
| cover\_url | object | 0.00% |

1. **Data Set Summary Statistics**

Narrative introduction to the section.

**Table 2: Summary Statistics for XXX (name of dataset)**

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| *Variable Name* | *Count* | *Mean* | *Standard Deviation* | *Min* | *25th* | *50th* | *75th* | *Max* |
| *Artist\_count* | *953* | *1.56* | *0.89* | *1* | *1* | *1* | *2* | *8* |
| *Released\_year* | *953* | *2018.24* | *11.12* | *1930* | *2020* | *2022* | *2022* | *2023* |
| *Released\_month* | *953* | *6.03* | *3.57* | *1* | *3* | *6* | *9* | *12* |
| *Released\_day* | *953* | *13.93* | *9.20* | *1* | *6* | *13* | *22* | *31* |
| *In\_spotify\_playlists* | *953* | *5200.12* | *7897.61* | *31* | *875* | *2224* | *5542* | *52898* |
| *In\_spotify\_charts* | *953* | *12.00* | *19.58* | *0* | *0* | *3* | *16* | *147* |
| *In\_apple\_playlists* | *953* | *67.81* | *86.44* | *0* | *13* | *34* | *88* | *672* |
| *In\_apple\_charts* | *953* | *51.91* | *50.63* | *0* | *7* | *38* | *87* | *275* |
| *In\_deezer\_charts* | *953* | *2.67* | *6.04* | *0* | *0* | *0* | *2* | *58* |
| *Bpm* | *953* | *122.54* | *28.06* | *65* | *100* | *121* | *140* | *206* |
| *Danceability\_%* | *953* | *66.97* | *14.63* | *23* | *57* | *69* | *78* | *96* |
| *Valence\_%* | *953* | *51.43* | *23.48* | *4* | *32* | *51* | *70* | *97* |
| *Energy\_%* | *953* | *64.28* | *16.55* | *9* | *53* | *66* | *77* | *97* |
| *Acousticness\_%* | *953* | *27.06* | *25.996* | *0* | *6* | *18* | *43* | *97* |
| *Instrumentalness\_%* | *953* | *1.58* | *8.41* | *0* | *0* | *0* | *0* | *91* |
| *Liveness\_%* | *953* | *18.21* | *13.71* | *3* | *10* | *12* | *24* | *97* |
| *Speechiness\_%* | *953* | *10.31* | *9.91* | *2* | *4* | *6* | *11* | *64* |

There should be a table for **EACH** categorical variable.

Table 3: Proportions for XXX (n=yyy)

|  |  |  |
| --- | --- | --- |
| *Category (artist\_count)* | *Frequency* | *Proportion (%)* |
| *1* | *587* | *61.59* |
| *2* | *254* | *26.65* |
| *3* | *85* | *8.92* |
| *4* | *15* | *1.57* |
| *5* | *5* | *0.52* |
| *6* | *3* | *0.31* |
| *7* | *2* | *0.21* |
| *8* | *2* | *0.21* |

After you summarize the categorical variables, generate a correlation matrix for all continuous variables (not categorical – this doesn’t make sense)

Table 4: Correlation Table/Tables

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **Artists\_count** | **Released\_year** | **Released\_month** | **Released\_day** | **In\_spotify\_playlists** |
| **Artist\_count** | 1.000000 | 0.087586 | 0.037995 | -0.017549 | -0.101966 |
| **Released\_year** | 0.087586 | 1.000000 | 0.076714 | 0.174282 | -0.392204 |
| **Released\_month** | 0.037995 | 0.076714 | 1.000000 | 0.080313 | -0.104757 |
| **Released\_day** | -0.017549 | 0.174282 | 0.080313 | 1.000000 | -0.079669 |
| **In\_spotify\_playlists** | -0.101966 | -0.392204 | -0.104757 | -0.79669 | 1.000000 |

After the table with the raw data, include a heatmap of the correlation matrix as a figure.

A screenshot of a computer screen

Description automatically generated

1. **DATA SET GRAPHICAL EXPLORATION**

Narrative introduction to the section. In each section below, indicate any interesting distributions, anomalies, imbalance, etc. that you notice.

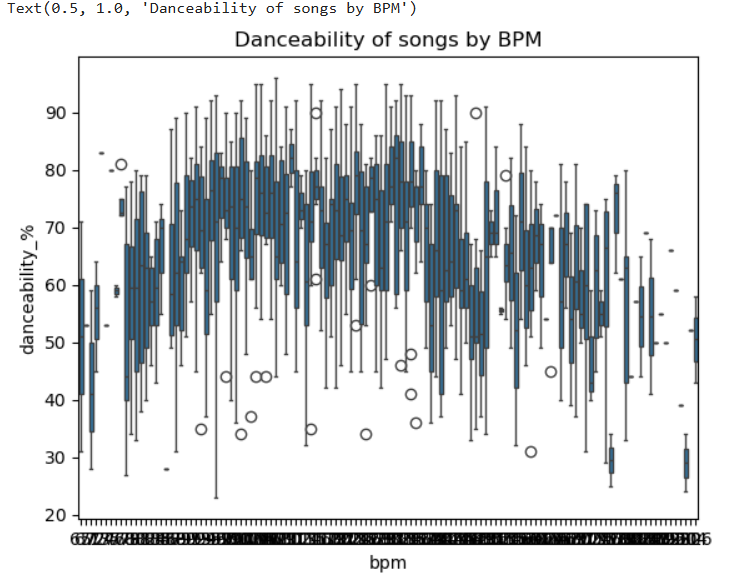
* 1. *Distributions*
  2. *ScatterPlots / Pairwise Plots (continuous variables)*
  3. *Barcharts (categorical variables)*
  4. *Other Plots - don’t skimp – there are likely other plots that would be useful that I haven’t already specified. Include those in this section.*

All figures should be cited formatted like this and mentioned in the text.

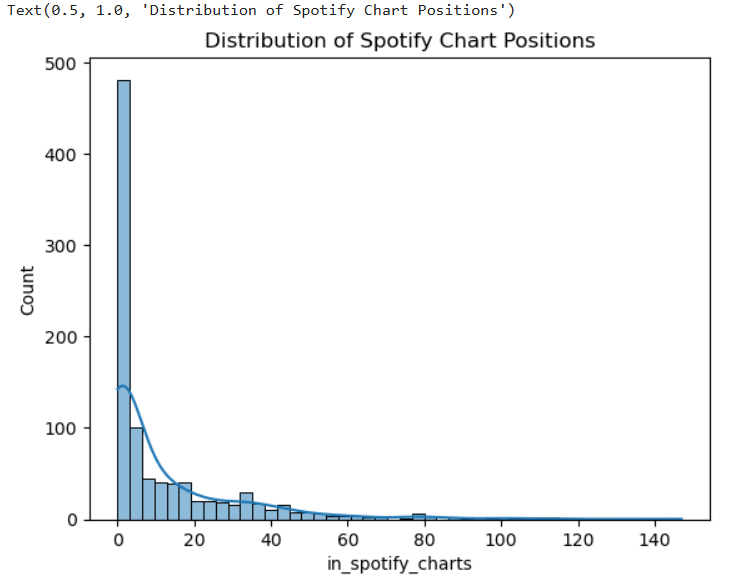
A chart with blue dots

Description automatically generated

**Figure 1: Songs by BPM from dataset (single plot)**

**

**Figure 2: Danceability of songs by BPM**

****

**Figure 3: Distribution of Spotify Chart Positions**

*A graph with blue lines

Description automatically generated*

**Figure 4: Number of Songs Released Each Month**

**A graph of a number of songs

Description automatically generated**

**Figure 5: Number of Songs Released Per Year**

1. **SUMMARY OF FINDINGS**

When analyzing this data set, we found that majority of songs that were streamed the most were released at the start of the year in January and early in the summer in May and June. Additionally, there was an even spread of data in our scatter plot regarding the number of songs and the songs’ beats per minute. We found that soon after 2020 there was a significant increase in the number of songs that were in released that were in Spotify’s most streamed songs. There was a decrease in the number of songs from 2023 because of this data set being published before the end of the year. Finally, we found that there is a strong relationship between the number of songs in the most streamed data set and their rank in the chart where 0 was the most popular and 147 was the least. This showed us that the most popular songs were also in the greatest number of playlists on Spotify.