**Spotify 2023**

**Exploratory Analysis (Final)**

Jonathan Penaloza ([jpenaloza@bellarmine.edu](mailto:jpenaloza@bellarmine.edu))

1. **INTRODUCTION**

For my final project in this class, I am planning to predict seasonal music trends given different musical factors and music platforms. Additionally, I will be identifying, as much as possible, which music platform is most relevant is the music market. Through this analysis, I hope to understand the preferences and behaviors of listeners during different seasons by analyzing data from various music platforms. The analysis aims to understand how factors like weather, holidays, and cultural events influence the popularity of certain genres and songs. The research will contribute to a better understanding of seasonal music trends and provide valuable information for musicians, platforms, and marketers. It may also shed light on the cultural and emotional associations people have with different seasons, enhancing our understanding of how music influences our moods and experiences. The independent variable for this dataset is the years from which these top 10 seasonal songs will be released from. The dependent variables are going to be the in\_spotify\_playlist and in\_spotify\_charts vs the in\_apple\_playlist and in\_apple\_charts. As well as the keys, modes, and bpm, (general number of) streams, release\_date. The various factors of interest of the public such as danceability\_%, valence\_%, energy\_%, acousticness\_%, instrumentalness\_%, liveness\_%, speechiness\_%. The dataset is mostly an integer data type with a few exceptions on the date, name of the songs, the key, and the mode. The data set contains a total of 25 columns and 954 columns. There are a few missing values, to which I would have to eliminate the rows since it is asking for things, I cannot just add using statistics. It is quite uncertain to know if the dataset is balanced, but I will still be using a logistic regression model and running multiple experiments to analyze the factors and characteristics of the top 10 songs of the season. I am planning to only use python for this project. I do think I will try to add more information to this project as I think my hypothesis for the project is going to require some background on the chosen season.

1. **DATA SET DESCRIPTION**

There are 953 entries (rows) in the dataset, with 25 columns. The columns position, name, their datatypes, number of, and percentage of missing values are as follows in **Table 1.**

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

|  |  |  |
| --- | --- | --- |
| *Variable Name* | *Data Type* | *Missing Data (%)* |
| *track\_name* | *Nominal, object* | *0%* |
| *artist(s)\_name* | *Nominal, object* | *0%* |
| *artist\_count* | *Interval, int64* | *0%* |
| *released\_year* | *Interval, int64* | *0%* |
| *released\_month* | *Interval, int64* | *0%* |
| *released\_day* | *Interval, int64* | *0%* |
| *in\_spotify\_playlists* | *Ordinal, int64* | *0%* |
| *in\_spotify\_charts* | *Interval, int64* | *0%* |
| *streams* | *Ordinal, object* | *0%* |
| *in\_apple\_playlists* | *Ordinal, int64* | *0%* |
| *in\_apple\_charts* | *Interval, int64* | *0%* |
| *in\_deezer\_playlists* | *Ordinal, object* | *0%* |
| *in\_deezer\_charts* | *Interval, int64* | *0%* |
| *in\_shazam\_charts* | *Ordinal, object* | *5.25%* |
| *bpm* | *Ordinal, int64* | *0%* |
| *key* | *Nominal, object* | *10%* |
| *mode* | *Nominal, object* | *0%* |
| *danceability\_%* | *Ratio, int64* | *0%* |
| *valence\_%* | *Ratio, int64* | *0%* |
| *energy\_%* | *Ratio, int64* | *0%* |
| *acousticness\_%* | *Ratio, int64* | *0%* |
| *instrumentalness\_%* | *Ratio, int64* | *0%* |
| *liveness\_%* | *Ratio, int64* | *0%* |
| *speechiness\_%* | *Ratio, int64* | *0%* |
| *release\_date* | *datetime64[ns]* | *0%* |

1. **Data Set Summary Statistics**

Since there are several columns in our data that do not contain data useful (or possible) to graph, these will be excluded, giving us 13 variables to examine the statistics for. The various statistics can be referenced in **Table 2**.

**Table 2: Summary Statistics for Spotify 2023**

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| *Variable Name* | *Count* | *Mean* | *Standard Deviation* | *Min* | *25th* | *50th* | *75th* | | *Max* | |
| *in\_spotify\_playlists* | *953* | *5200.125* | *7897.6* | *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* | *51.90* | *50.63* | *0* | *7* | *38* | *87* | | *275* | |
| *in\_deezer\_charts* | *953* | *2.70* | *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* | *25.99* | *0* | *6* | *18* | *43* | | *97* | |
| *instrumentalness\_%* | *953* | *1.58* | *8.40* | *0* | *0* | *0* | *0* | | *91* | |
| *liveness\_%* | *953* | *18.21* | *13.71* | *3* | *10* | *12* | *24* | | *97* | |
| *speechiness\_%* | *953* | *10.13* | *9.91* | *2* | *4* | *6* | *11* | | *64* | |
| *release\_date* | *153* | *2018-01-06 13:29:24.705882368* | *NaN* | *1946-11-01 00:00:00* | *2019-12-06 00:00:00* | *2021-12-24 00:00:00* | | *2022-12-02 00:00:00* | | *2022-12-30 00:00:00* |

The following tables **Table 3** and **Table 4**, show the categorical variables within the dataset.

Table 3: Proportions for Mode Type

|  |  |  |
| --- | --- | --- |
| *Mode* | *Frequency* | *Proportion (%)* |
| *Major* | *550* | *57.62%* |
| *Minor* | *403* | *42.38%* |

**Table 4: Proportions for Type of Key**

|  |  |  |
| --- | --- | --- |
| *Key* | *Frequency* | *Proportion (%)* |
| *C#* | *215* | *22.5%* |
| *G* | *96* | *10.07%* |
| *G#* | *91* | *9.5%* |
| *F* | *89* | *9.33%* |
| *B* | *81* | *8.49%* |
| *D* | *81* | *8.49%* |
| *A* | *75* | *7.86%* |
| *F#* | *73* | *7.66%* |
| *E* | *62* | *6.50%* |
| *A#* | *57* | *5.98%* |
| *D#* | *33* | *3.46%* |

The following figure is a correlation matrix for all the continuous variables within the dataset.

A green and black squares with numbers

Description automatically generated with medium confidence

**Figure 1: Correlation matrix pf continuous variables excluding categorical variables.**

Based on the heatmap both population in Spotify and Deezer seem to have a more similar likings for the same songs than the population of apple users, but all together seem to like the same songs. Ergo there is a higher correlation of Deezer and Spotify users that listen to the same songs than the apple users.

Table 5: Correlation Table/Tables

The following table displays the songs within the dataset that had the highest percentage of energy.

|  |  |  |
| --- | --- | --- |
| ***Track Name*** | ***Artist Name*** | ***Energy %*** |
| *I'm Good (Blue)* | *Bebe Rexha, David Guetta* | *97* |
| *Murder In My Mind* | *Kordhell* | *97* |
| *That That (prod. & feat. SUGA of BTS)* | *PSY, Suga* | *96* |
| *Tá OK* | *dennis, MC Kevin o Chris* | *96* |
| *Bombonzinho - Ao Vivo* | *Israel & Rodolffo, Ana Castela* | *95* |
| *Merry Christmas* | *Ed Sheeran, Elton John* | *94* |
| *Every Angel is Terrifying* | *The Weeknd* | *94* |
| *Idol (「アイドル」)* | *YOASOBI* | *94* |
| *KICK BACK* | *Kenshi Yonezu* | *94* |
| *Freaks* | *Surf Curse* | *94* |

Table 6: Correlation Table/Tables

The following table displays the songs within the dataset that had the lowest percentage of energy.

|  |  |  |
| --- | --- | --- |
| ***Track Name*** | ***Artist Name*** | ***Energy %*** |
| *Happier Than Ever* | *Billie Eilish* | *24* |
| *It's Beginning To Look A Lot Like Christmas* | *Michael Buble* | *23* |
| *Boyfriends* | *Harry Styles* | *20* |
| *Special* | *SZA* | *20* |
| *I'm Tired - From "Euphoria" An Original HBO Se...* | *Labrinth* | *20* |
| *Something In The Way - Remastered 2021* | *Nirvana* | *20* |
| *Sweet Nothing* | *Taylor Swift* | *16* |
| *The Christmas Song (Merry Christmas To You)* | *Nat King Cole* | *15* |
| *Heart To Heart* | *Mac DeMarco* | *14* |
| *What Was I Made For? [From The Motion Picture Barbie]* | *Billie Eilish* | *9* |

*When looking at which tracks have the most energy and which tracks have the least energy, the song "I'm Good Blue" by Bebe Rexha and David Guetta had the highest percentage of energy while the song "What Was I Made For? from the Motion Picture Barbie" had by Billie Eilish the least percentage of energy.*

1. **DATA SET GRAPHICAL EXPLORATION**

**A graph with blue and white bars

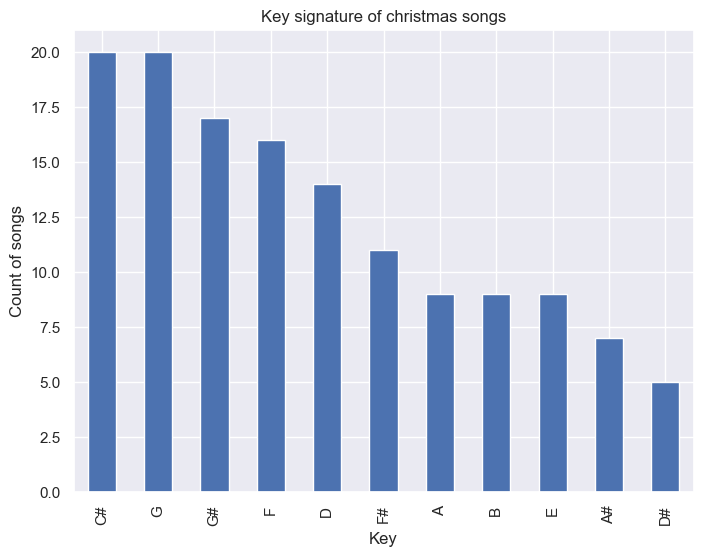
Description automatically generated**

**Figure 1. Returns the top 9 most listened songs based of the number of streams.**

*Based on this bar graph, in 2023, The artist with the most streams on Spotify was the SZA with 96709329, while the artist with the second most streams was The Walters with 972164968 streams, and the artists with the third most streams being Burna Boy.*

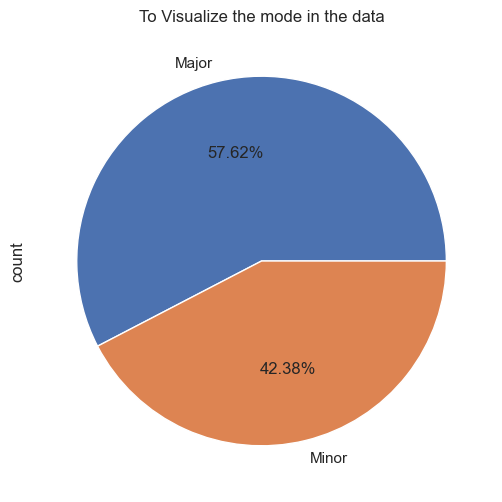
**Figure 2. Identify which key most of Christmas songs appear in the data.**

**The most used key are C# and G with 20 songs with the same key signature and the least used key is D# with only 5 songs.**

**

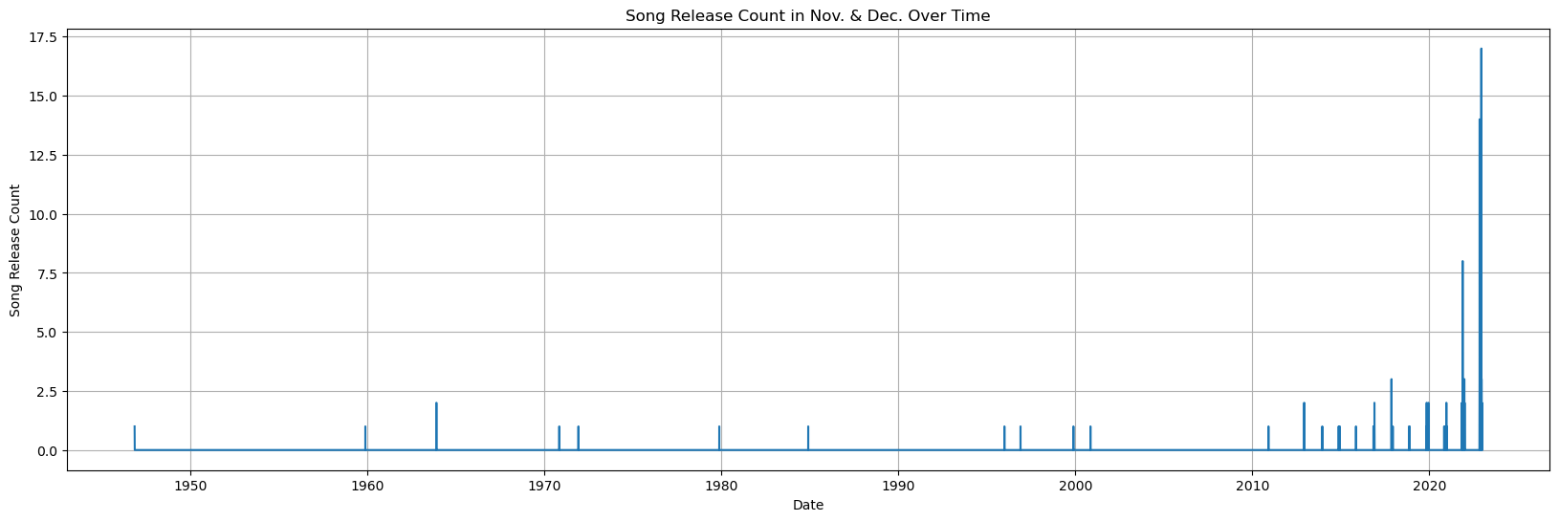
*Figure 3. Identify which mode is most used in Christmas songs.*

*57.62% of all the songs that are written in major mode while there is a 42.38% of songs written in minor mode.*

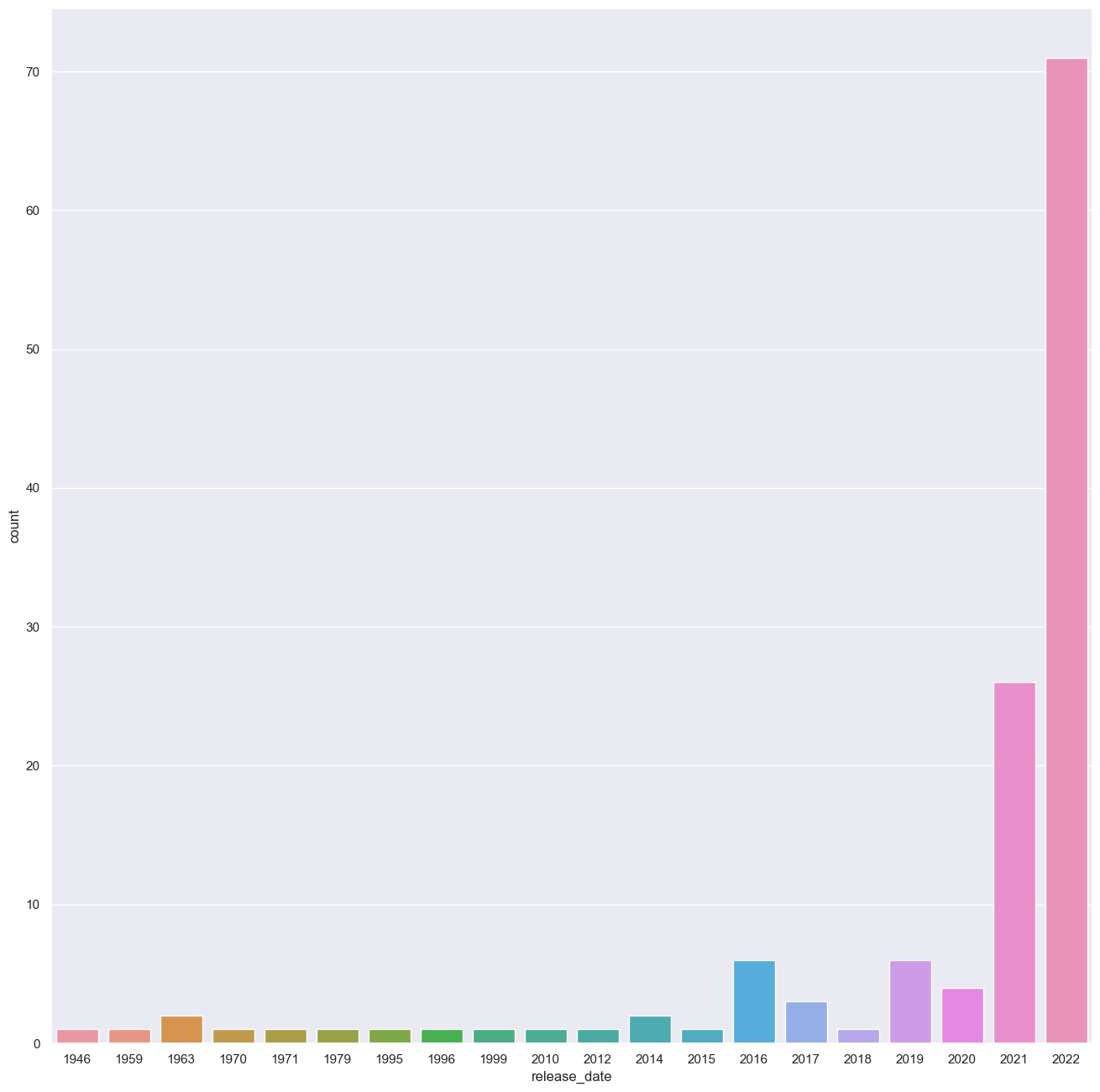
**

*Figure 4. Song Release Count in Nov. & Dec. Over Time*

*This shows the count of songs released in November and December over the years.*

**

*Figure 5. This is another way of representing the same data but with a bar graph.*

**

*Figure 6. This bar graph helps with understanding Christmas songs by the used keys.*

*Here we can see that over time we tended to have very similar keys lasting up to decades, until most recent years where we can see a great diversity of keys used to make Christmas music.* A graph with different colored lines

Description automatically generated

**Figure 7: Create histogram to understand Christmas music using mode.**

**Here the same idea as figure 6 is explored but comparing the mode, which shows a very similar tendency.** A graph with numbers and lines

Description automatically generated

Observations:

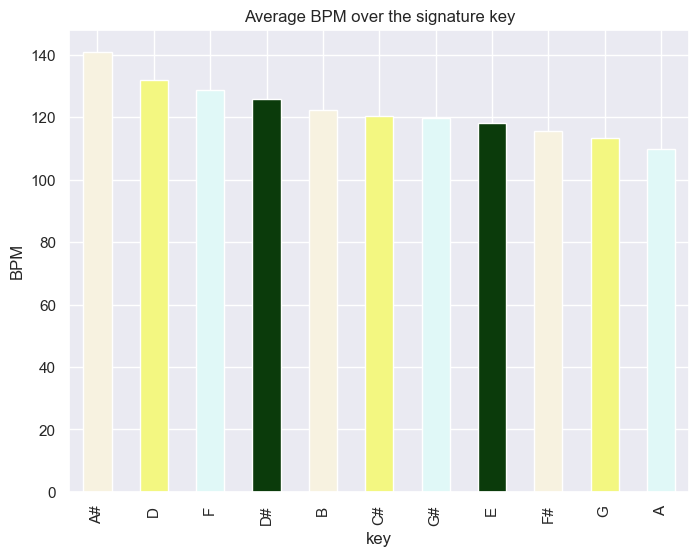
From the Two histograms we observed some interesting things about the diversity of keys and mode of the songs.

For the key, there is mode diversity of keys used in the newly released Christmas music.

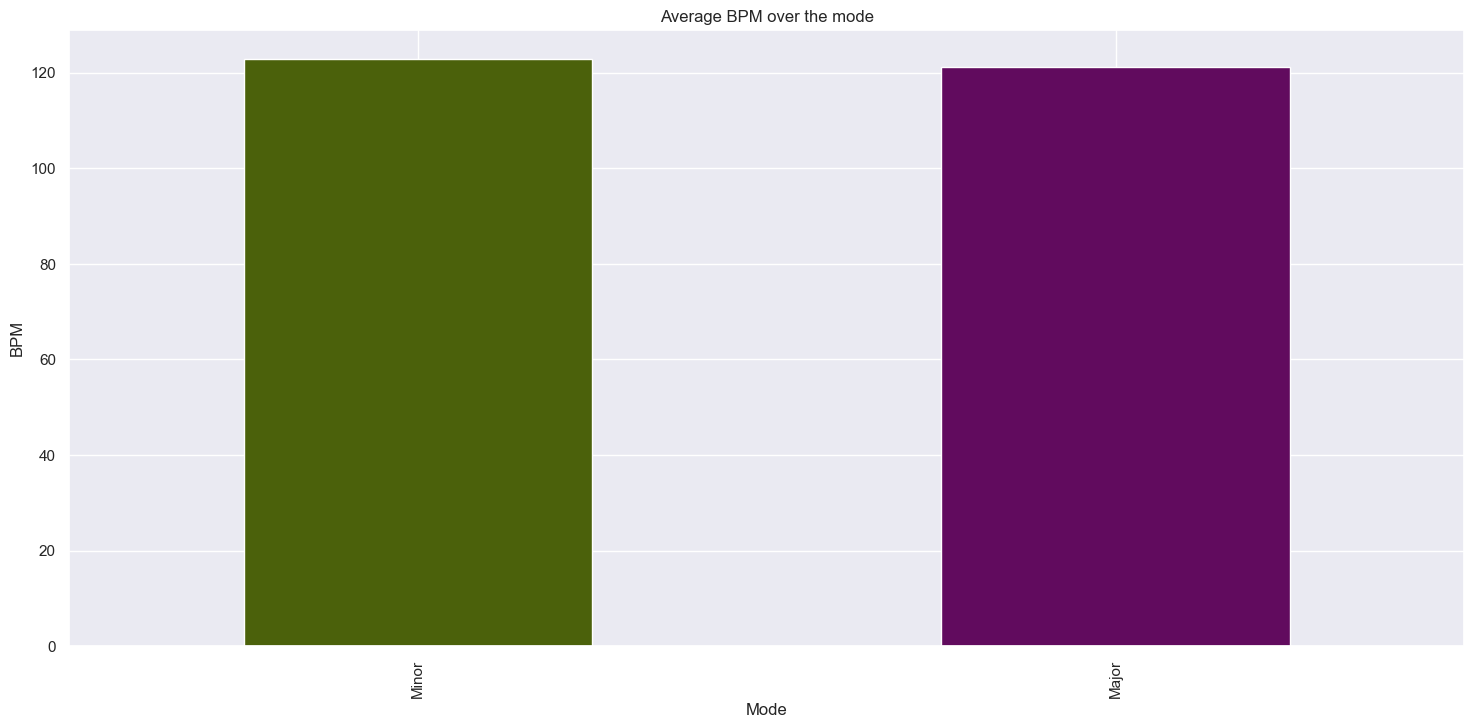
In the Second chart we visualize that there is a more use of both modes, even after having clear trends of modes over the years.

As predicted G#, C# and G are the most used key signatures.

**Figure 7: Visualize the bar chart the average BPM over the signature key.**



**Figure 8: Visualize the bar chart the average BPM over modes.**



*Observations:*

*From the first chart we can see that certain songs have a higher speed regardless of how popular those keys might be.*

*In the Second chart we can see that both major and minor modes have very similar bpm.*

*Figure 9. Linear Regression*

*From the use of the x and y variables for the linear regression after some trial and error we get this outcome:*

*x=ds[['released\_year', 'released\_month', 'released\_day', 'in\_spotify\_playlists','in\_spotify\_charts', 'in\_apple\_playlists', 'in\_apple\_charts','in\_deezer\_playlists', 'in\_deezer\_charts', 'in\_shazam\_charts','danceability\_%', 'valence\_%', 'energy\_%','acousticness\_%', 'instrumentalness\_%', 'liveness\_%', 'speechiness\_%’]]*

*y=ds[['bpm’]]*

*MSE: 812.27*

*RMSE: 28.50*

*R-square: -0.24*

*A screenshot of a computer code

Description automatically generated*

1. **SUMMARY OF FINDINGS**

Based on my findings, the year most Christmas songs were created was 2022, which makes sense given the newly given technological capabilities to create music. During that time, artists and musicians were able to explore new sounds and experiment with innovative production techniques. This led to a surge in the creation of Christmas songs as they embraced the opportunity to bring fresh and modern elements to this beloved genre. Which leads to the next point, the variability of key signatures and modes that we are being presented with. As the graphs have shown us, there have been tendencies over decades with the same signature key and mode, but as we approach our times, we can see that music is varying more than ever in terms of key signatures and modes. This can be attributed to the increasing diversity and globalization of music, as artists draw inspiration from different cultures and genres. Additionally, advancements in technology have made it easier for musicians to experiment with different keys and modes, resulting in a wider range of musical expressions.

There were not many changes with the number of beats per second, as it seems relatively normal, for both modes but not for all keys. Demonstrating that there are some keys that, regardless of how relevant they may be, seem to require a certain tempo to make sense of the idea of Christmas music. This observation suggests that the choice of tempo in Christmas music is not arbitrary and plays a crucial role in creating the desired festive atmosphere. It implies that certain keys may have inherent associations with the holiday season, requiring specific tempos to evoke the intended emotions and sentiments.

I really thought that there could be more said about the Spotify charts, Apple Charts, Shazam, and Deezer, but by seeing the heat map and other statistics, they seem relatively like each other. These platforms, including Spotify, Apple Music, Shazam, and Deezer, share similarities in terms of their heat maps and statistics. However, it is important to note that, while they may appear similar on the surface, each platform has its own unique features and user base. Exploring these differences can provide a deeper understanding of how music consumption varies across different streaming services.