Spotify

Exploratory Analysis

Reese, rbottorff#@bellarmine.edu

Michael, mzelaya@bellarmine.edu

I. INTRODUCTION

Our dataset is about the stats of Spotify music from 1986 to 2023. This dataset contains 28 columns with information about Track, Album, Artist, and Audio. We chose this dataset because we both enjoy music and thought it would be interesting to explore and investigate.

II. DATA SET DESCRIPTION

- track_name: The name of the song.

- popularity: The song's popularity level on Spotify (popularity score).

- disc_number: The disc number to which the song belongs on an album.

- duration_ms: The duration of the song in milliseconds.

- explicit: Indicates whether the song contains explicit content (True or False).

- track_number: The track number of the song on the album.

- album_name: The name of the album.

- album_release_date: The release date of the album.

- album_type: The type of album (e.g., "album," "single," "compilation," etc.).

- album_total_tracks: The total number of songs on the album.

- artists_names: The names of the artists who perform the song (there can be multiple, separated by semicolons).

- principal_artist_name: The name of the principal artist.

- artist_genres: The music genres associated with the principal artist.

- principal_artist_followers: The number of followers of the principal artist on Spotify.

- acousticness: An indicator of the song's acoustic characteristics.

- danceability: An indicator of how danceable the song is.

- energy: The perceived energy of the song.

- instrumentalness: An indicator of the presence of instrumental elements in the song.

- key: The musical key of the song.

- liveness: An indicator of the likelihood that the song was performed live.

- loudness: The loudness of the song.

- mode: The musical mode of the song.

- speechiness: An indicator of the amount of speech in the song.

- tempo: The tempo of the song.

- time_signature: The time signature of the song.

- valence: A measure of the positivity of the song.

- year: The year in which the song was released. It is of integer type (int64).

- duration_min: The duration of the song in minutes, rather than milliseconds. It is of float type (float64).

Table 1: Data Types and Missing Data

Variable Name	*Data Type*	*Missing Data (%)*
Track_name	Object	0%
Popularity	Int	0%
Disc_number	Int	0%
Duration_ms	Int	0%
explicit	Bool	0%
Track_number	Int	0%
Album_name	Object	0%
Album_release_date	Object	0%
Album_type	Object	0%
Album_total_tracks	Int	0%

Artist_names	Object	0%
Principal_artist_name	Object	0%
Artist_genres	Object	0%
Principal_artist_followers	Int	0%
Acousticness	Float	0%
Danceability	Float	0%
Energy	Float	0%
Instrumentalness	Float	0%
Key	Float	0%
Liveness	Float	0%
Loudness	Float	0%
Mode	Float	0%
Speechiness	Float	0%
Tempo	Float	0%
Time_signature	Float	0%
Valence	Float	0%
Year	Int	0%
Duration_min	Float	0%

III. 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*	*25%*	*50%*	*75%*	*Max*

Popularity	11450.00	67.999301	9.334255	44.00	61.00	68.00	75.00	100.00
Disc_number	11450.00	1.017380	0.195043	1.0	1.0	1.0	1.0	10.0
Duration_ms	1.145	2.3018	1.1812	3.3493	1.9172	2.2321	2.5911	1.0828
explicit	Bool	0%						
Track_number	11450.00	5.2517	4.4159	1.0	2.0	4.0	8.0	48.0
Album_total_tracks	11450.00	14.2065	9.5481	1.0	11.0	13.0	16.0	176.0
Principal_artist_followers	1.145	9.6900	1.8263	3.10	9.0695	2.9017	8.6264	1.1467
Acousticness	11450.00	0.2256	0.2614	0.0	0.0183	0.1120	0.3620	0.9960
Danceability	11450.00	0.6109	0.1613	0.0	0.5070	0.6210	0.7280	0.9880
Key	11450.00	5.2675	3.5436	0.0	2.0	5.0	8.0	11.0
Liveness	11450.00	0.1829	0.1450	0.0	0.9220	0.1250	0.2370	0.9820
Loudness	11450.00	-7.5334	3.7758	-47.0700	-9.2077	-6.7195	-5.0210	0.5220
Mode	11450.00	0.6778	0.4673	0.0	0.0	1.0	1.0	1.0
Speechiness	11450.0	0.0875	0.0929	0.0	0.033	0.0473	0.0936	0.9440

	0				9			
Tempo	11450.00	121.015	30.182	0.0	96.94	119.210	140.058	220.099
Time_signature	11450.00	3.93	0.35	0.0	4.0	4.0	4.0	5.0
Valence	11450.00	0.538	0.245	0.0	0.346	0.541	0.736	0.994
Year	11450.00	2004	11.03	1986	1995	2004	2014	2023
Duration_min	11450.00	3.83	1.96	0.55	3.19	3.720	4.31	180.469

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

Table 3: Proportions for XXX (n=yyy)

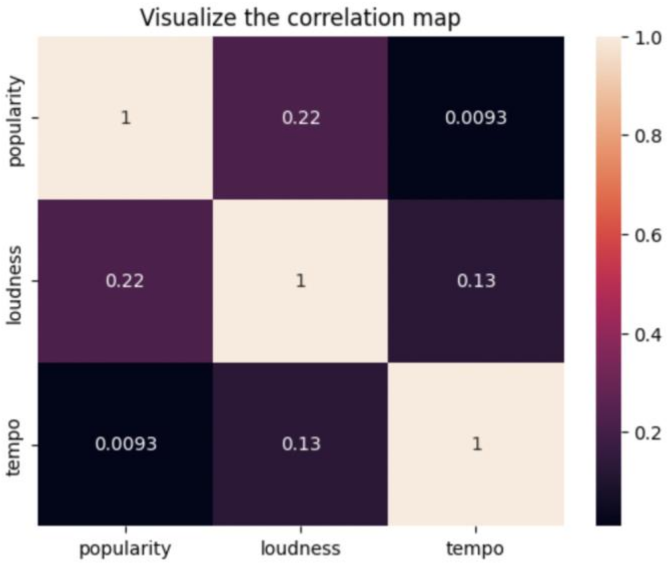
Category(album_type)	*Frequency(acousticness)*
Album	9815
Compilation	708
Single	927

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

Album_type	*Instrumentalness*
Single	0.053335
Album	0.04566
Compilation	0.31101

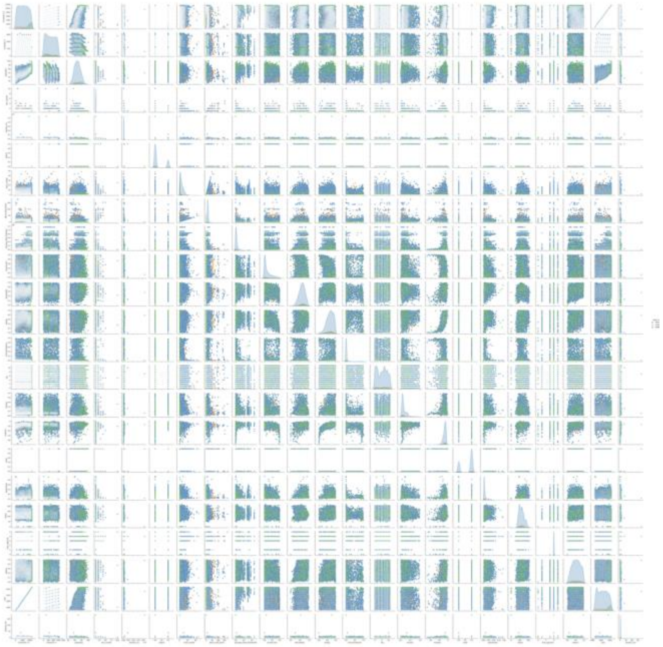
Correlation for Popularity



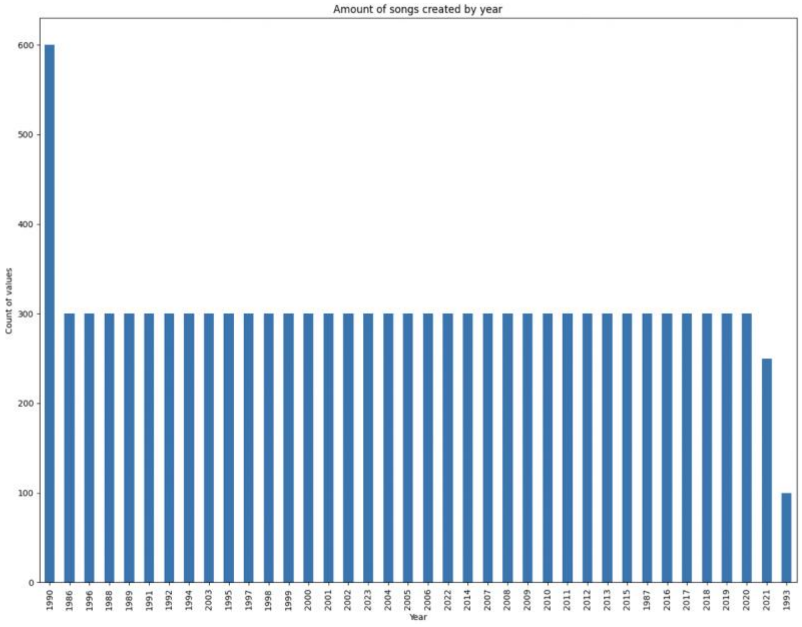
After the table with the raw data, include a heatmap of the correlation matrix as a figure.

IV. DATA SET GRAPHICAL EXPLORATION

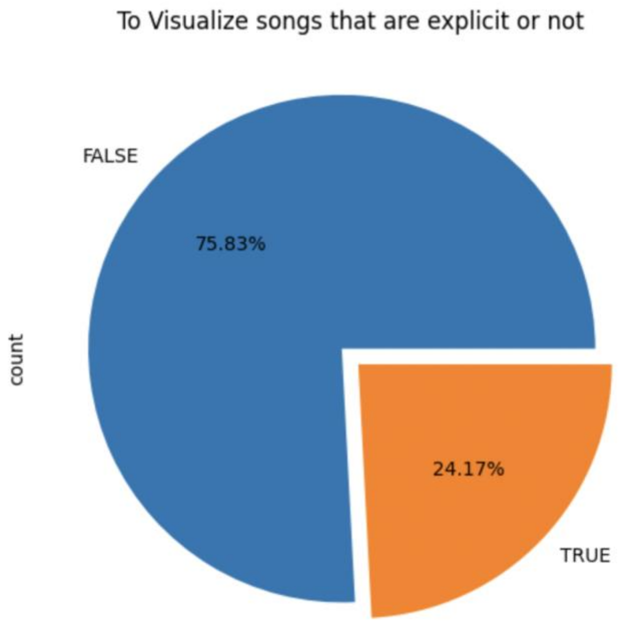
Narrative introduction to the section. In each section below, indicate any interesting distributions, anomalies, imbalance, etc. that you notice.



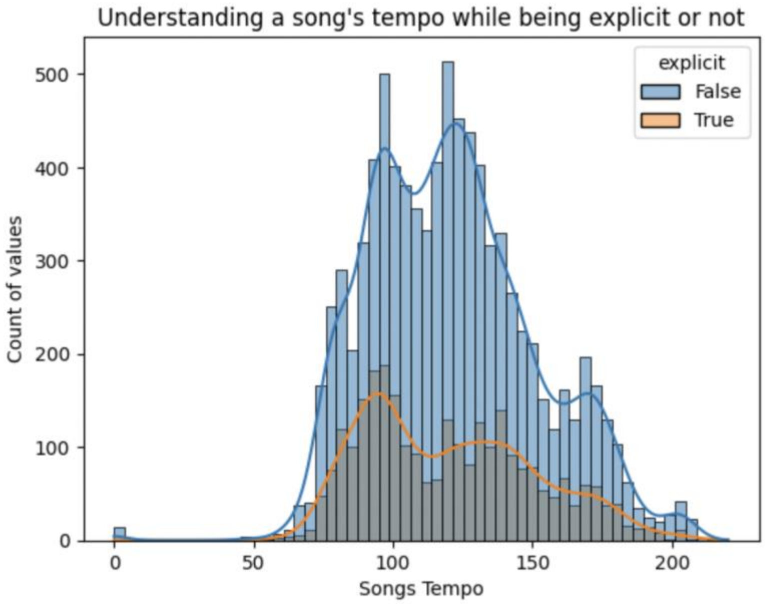
- This visualization shows that most songs are albums, a good amount is compilation, and a few are singles from 1986 to 2023



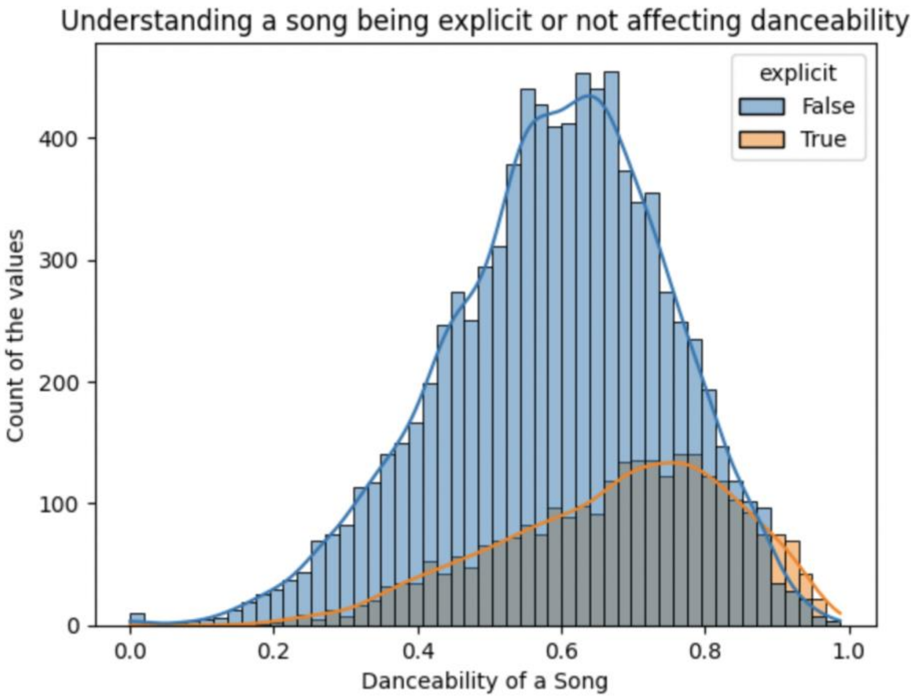
- From this observation, 1990 had the most songs (600), while 1993 had the least amount (100). Additionally, the other years in the data set had the same amount of 300 except for 2021, with 250 songs created



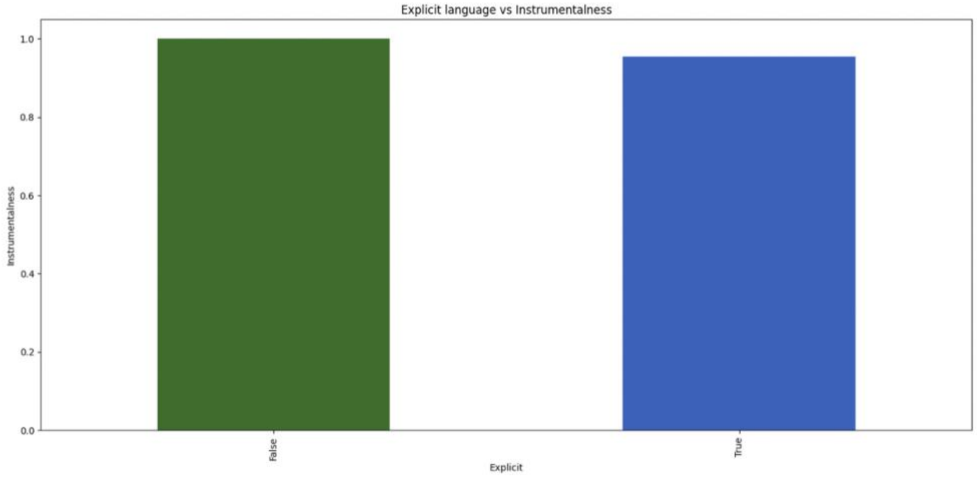
- From this visualization, a good number of songs are non-explicit (76%), while there are a few explicit songs (24%) in this data set



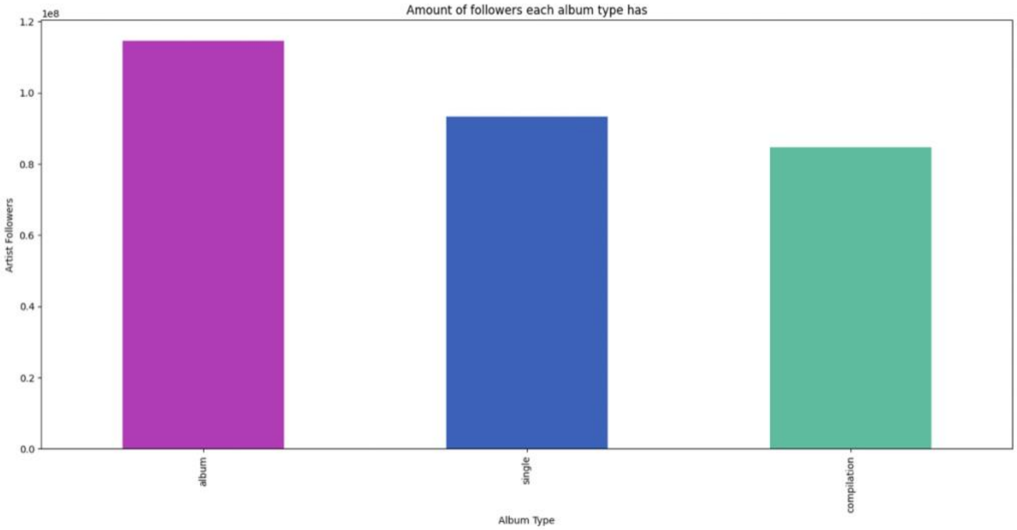
- This observation demonstrates that songs without being explicit had a higher tempo than songs with inappropriate music



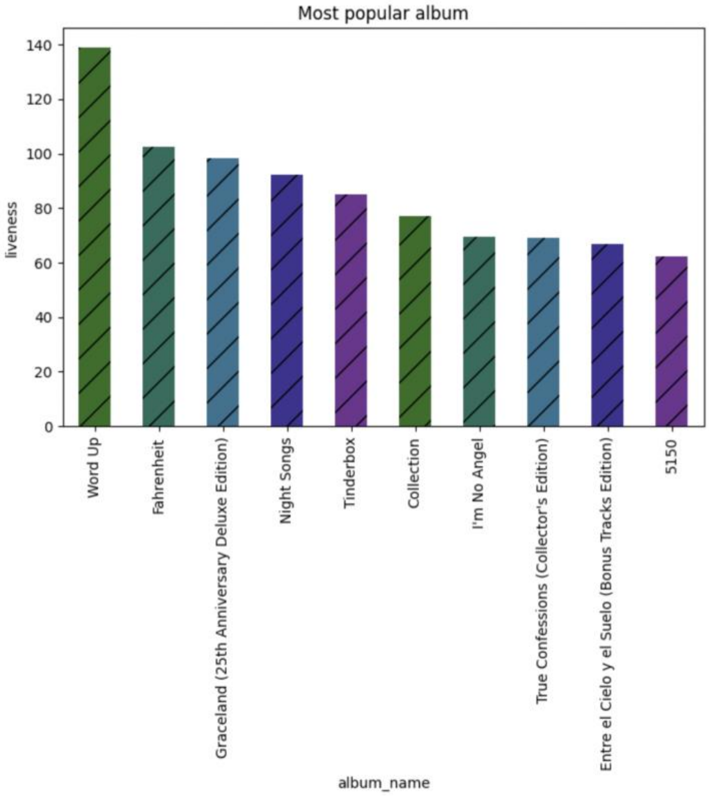
- From this observation, songs with explicit music had a higher danceability measurement than songs being appropriate



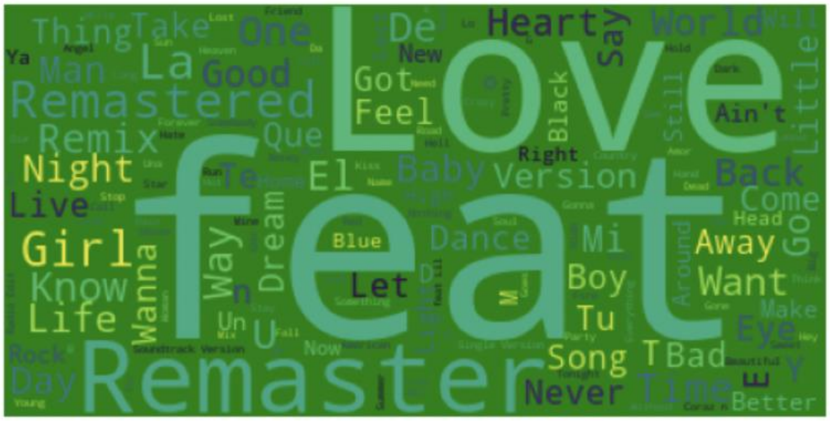
- The bar graph shows that the songs with more instruments has less explicit words than the songs with less instruments



- The bar graph shows that albums have the most followers and compilations have the least amount of followers. This can also tell us that albums are the most popular pick for people



- This bar graph shows the 10 most popular albums out of the entire dataset. This shows the relationship between the song and the liveness of the song. Word Up has the highest liveness which means there's a lot of energy which gains them a lot of followers



- This Worldcloud graph shows all the words that were used throughout this dataset, the most frequent ones being the largest in size

V. SUMMARY OF FINDINGS

The dataset chosen, "Spotify", shows the statistics of music from 1986-2023. In this dataset, there are 28 different variables being used, ranging from album type, to the danceability of the music. This dataset allowed us to dive deep into comparing variables and finding interesting information. Using the describe() function, we were able to see many things. One thing that stood out to us was that the mean year that music was released was 2004. We also found it interesting that the max minutes was 180, showing that there was a song that went on for 180 minutes.

Looking further into the dataset, we created graphs to show the different explorations we did. We were able to identify the year with the most amount of songs released, using a bar graph, showing that 1990 had the highest release year with 600 songs. We used a pie chart to show the percentage of songs that contained explicit language, showing about 75% of songs did not contain explicit language. We also used a hist plot to see if the explicit music had a higher danceability rate, and according to the graph it has a higher rate. We used bar graphs to see the relationship between album types and the amount of followers, showing that albums had the most followers whereas compilations had the least amount of followers. We used another bar graph to show the 10 most popular songs out of the entire dataset. We then created a proportion table for album types to see how frequent the specific types created acoustic music. This showed us that albums had the most acousticness and compilations had the least amount of acousticness.

Overall, this dataset was very interesting to explore and investigate. We were shocked with some of our findings and enjoyed learning more about the staticstical side of music. It really opened our eyes seeing how deeply people look into that statistic part of music. Completing explorations like ours can help producers create music that pleases everyone.