Relationship between Musical Features and Spotify Streams

Yining Wang (yw2622)

Chenyu Yu (cy523)

Yiwei Luo (yl3928)

a) Screenshot for Project 1:

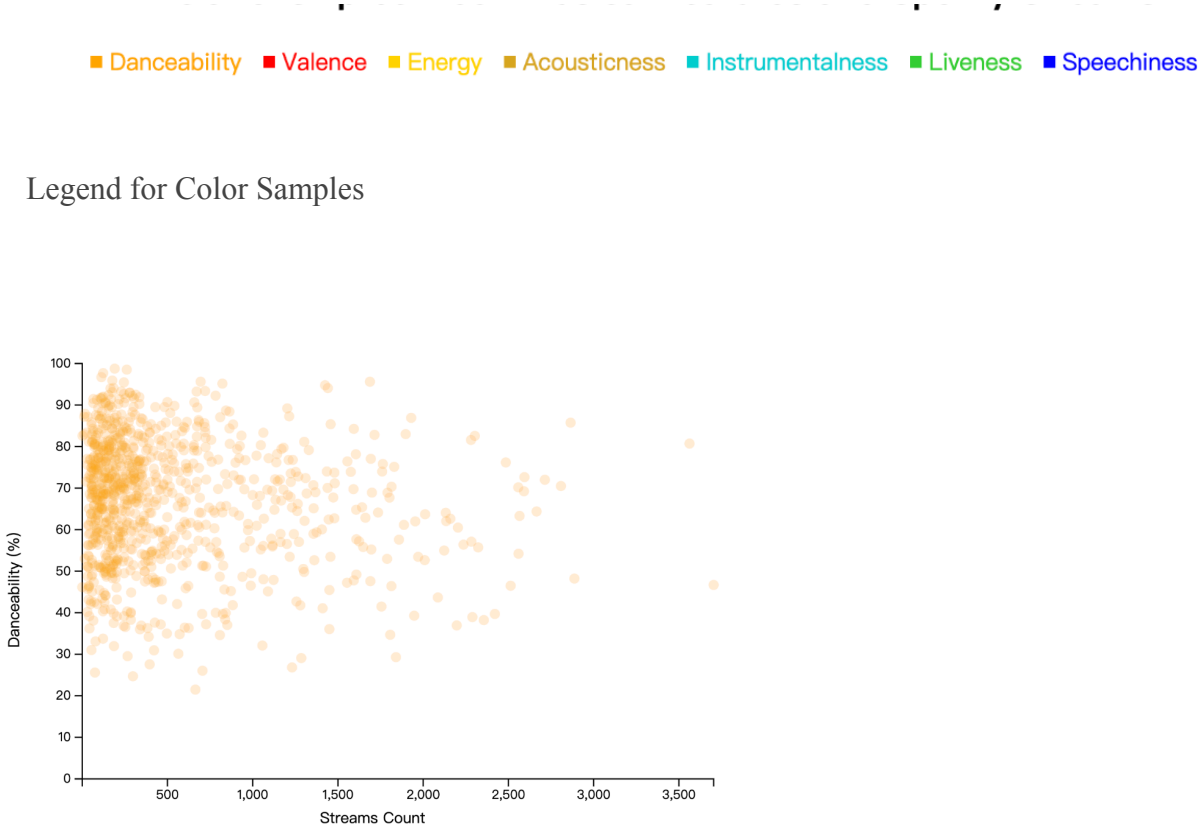

Figure 1. Streams Count vs. Danceability Scatter Plot

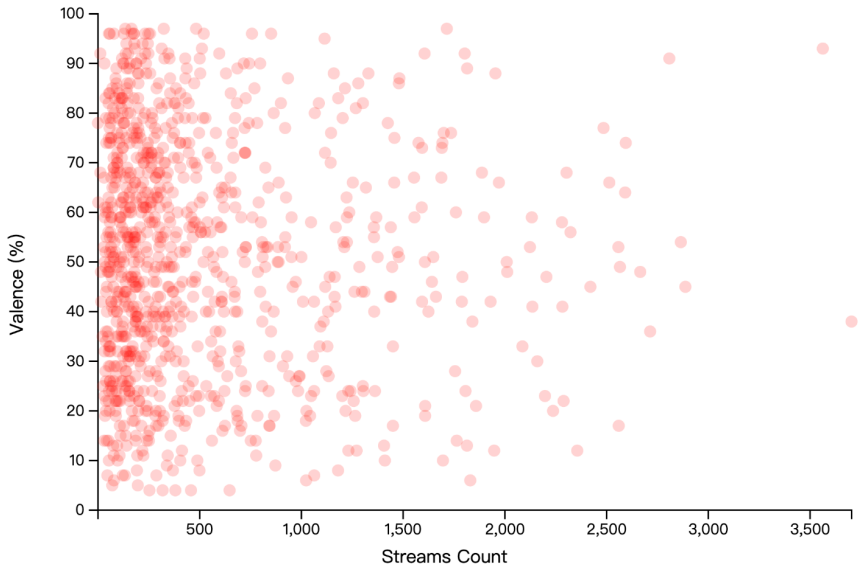

Figure 2. Streams Count vs. Valence (%) Scatter Plot

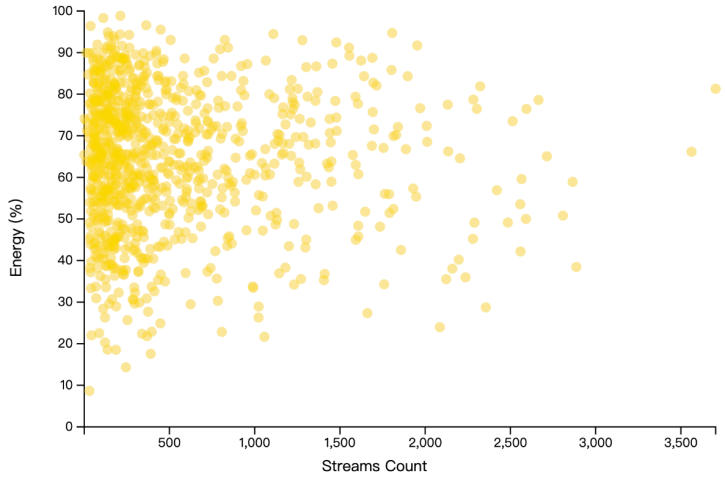

Figure 3. Streams Count vs. Energy (%) Scatter Plot

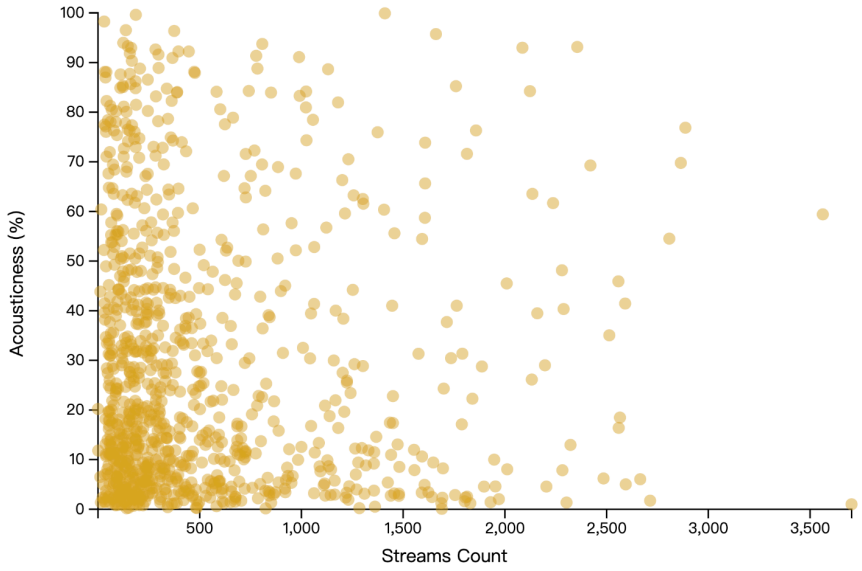

Figure 4. Streams Count vs. Accousticness (%) Scatter Plot

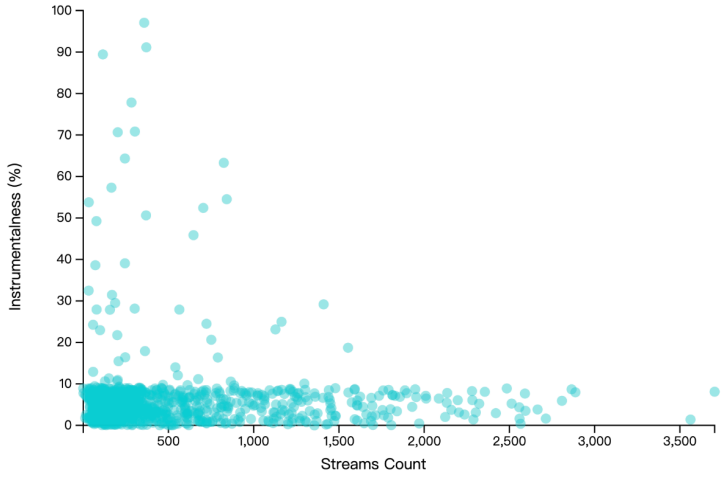

Figure 5. Streams Count vs. Instrumentalness(%) Scatter Plot

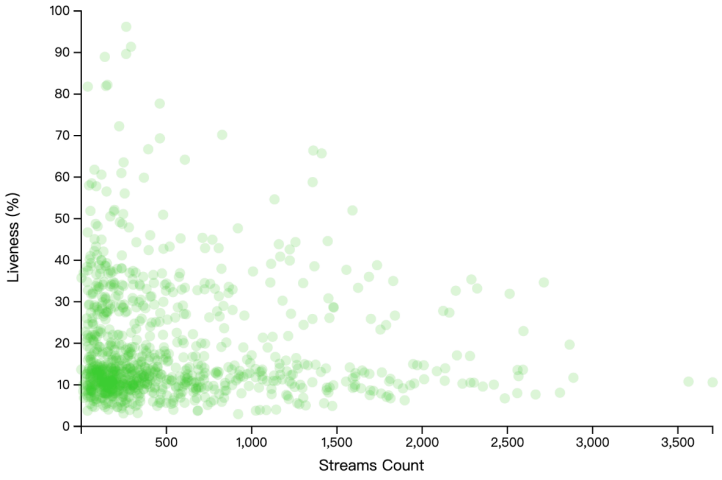

Figure 6. Streams Count vs. Liveness (%) Scatter Plot

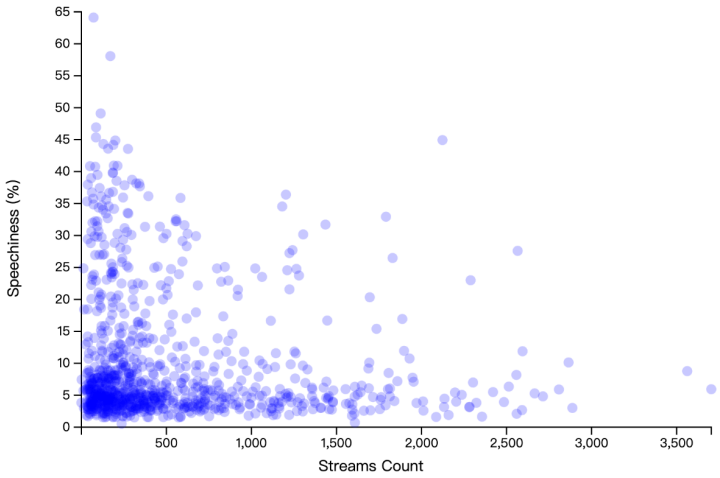

Figure 7. Streams Count vs. Speechiness (%) Scatter Plot

b) Data Description

The team extracted the data from a popular data science practice website named Kaggle. The link to the website can be found in the citation at the end of the report. It is a

dataset about the comprehensive information on some of the most streamed songs on music streaming platform Spotify supplemented with additional insights from other music streaming platforms like Apple Music. The basic data include track name, artist name, artist count, and release day, month and year. The relevant seven subsets of data for the report includes suitability for the song for dancing as “danceability_%”, positivity of the music content as “valence_%”, perceived energy level of the song as “energy_%”, acoustic song presence level as “acousticness_%”, proportion of instrumental content as “instrumentalness_%”, presence of live performance element as “liveness_%”, and amount of speech word in music as “speechiness_%” **(**Abdullah, 2024). Those data parameters are extracted from the given data set and put into a scale with scaleLinear() function as values in y axis. They are plotted with one single sub-dataset on x axis, parameter “streams”, which represents the total number of streams on Spotify in 10 million as parameter "streams(10 million)”, to plot the seven respective comparative scatterplot with the seven aforementioned data sets.

In convenience to the development, all of the values from the seven datasets are chosen because they contain only numbers in the form of strings. There are no exceptions that might cause type errors such as characters ($. %, and &) and word strings. Also, they are characteristics of a song that are commonly perceived to be related to the popularity of songs, which is represented by number of stream counts in the “streams” parameter.

To reformat the string data, the team utilized forEach loop for the data and the type coercion of JavaScript language to directly append, using JavaScript version of operation of “+=” to append the data into a new dataset with simplified name. One special case is the parameter “streams(10 million)”. Because the original data values are small in value,

as it was counted in the unit of 10 million, it is multiplied by 10 when appended into the "stream" parameter to avoid cluttering and recount the values in the more common unit of millions. To further avoid the visual representation of data point cluttering, a jitter function, which returns a random number from a given minimum and maximum number, is employed and added to the seven y-axis values. The minimum value of the jitter function is set to a negative number for scales with data points lowest in values perceptibly higher than the x axis while excluding those with data points directly on the x-axis to avoid the points to go beneath the scale. Similarly, depending on the visual representation of the location of data clustering on y scale and the distance between the end of scale of y-axis and the largest data in the two axes, a positive variable max is set to be the maximum value to be returned in the jitter function to avoid both cluttering and overscaling issues.

c) Overview of Design Rationale

To begin with, a specific division with "Legend" class is positioned directly under the group member names and title and above the graphs at the center of the screen. It leaves 20 pixels of margin between the title and the plots. Each "legend" item within the class is presided side by side. The color representing each of the seven y-axis variables is designated in a small square of 8 x 8 squared pixel size followed by the name of the variable in the same color that represents it, padded 5 pixels on its right. All of the samples are designated to be positioned in the center with 10 pixels between them.

The team chose scatterplot for the design because it is easier for the audience and developers to identify the pattern in scatterplot through the clustering visual representation of data points. It shows how data of the seven attributes are distributed in comparison to the popularity

demonstrated by x-values of stream counts. Also, the data of the attributes from the most popular song is not intended by entry to be in a continuous relationship demonstrated by a path. It is more suited for scatterplot to demonstrate the data in a cluster. Last but not least, the team believed that the project aims to present qualitative data showing a pattern. Therefore, the cluster pattern demonstrated by scatterplot is key for the audience to understand the concentration of popularity in relation to the seven attributes.

The team chose a svg canvas with 600 pixels in width and 400 pixels in height to demonstrate the plots in a more visually appealing and acceptable size to the audience. The top, 20 pixels, and bottom, 50 pixels, of padding is provided to separate the plots from each other top and down with 70 pixels of space to avoid clustering of plot graphs. The 70 pixels of left padding and 30 pixels of right padding position the plot in the center of the screen when scrolled for a more visually pleasing experience. Texts that explain the seven attributes on the y axis and the stream count value on the x axis are positioned through text-anchor in the middle of both x-axis and y-axis to avoid being written out of the canvas and suit the norm of similar visual plots.

Each scatterplot is filled with different colors that are conventionally accepted metaphor to the y-axis data they represent. For instance, orange is assigned to danceability because it is associated with enthusiasm, happiness, creativity related to the action of dancing. Red is assigned to valence, the positivity of the song, as it is related to love, romance, and power which is related to positivity. Gold is assigned to energy as it is associated with warmth and fortune. Goldenrod is assigned to acousticness because it is the color of woods, which one of the most known acoustic instrument guitar is made out of. Dark turquoise is used to instrumentalness as it represents professionalism and objectivity, which a musical instrument is intended to be played. Limegreen is chosen to represent liveness because it is commonly associated with plants and life. Blue is

chosen to represent speechiness as it is utilized in messenger to represent text messages. Also, the distinct colors aim to distinguish each scatterplot and data set on y axis it represents with each other. All of the colors are generally positioned in a bright spectrum of hue and color saturation to demonstrate the youthful, energetic nature of the majority of the younger demographic of listeners of Spotify and popular music.

Due to the large amount of overlapping between data points, the opacity of the circles representing the data points are tuned to 0.2 and 0.5 to reduce the amount of overlap and demonstrate the pattern of concentration of data points. For colors on the brighter spectrum of color hues, their opacity are tuned higher into 0.5 to make it visible against the white background. Due to the massive amount of data present, each data point is attributed to circle dots with a relatively smaller size of 4 pixels. Dot is chosen to be the shape of data points because of its generalized association to a data point. No jitter function is added to the valence chart because the overlapping of data points is relatively not severe. For acousticness and instrumentalness scales, while max value is set in a positive value, the min value is set to be zero to avoid their lowest data points perceptibly adhered on the x axis to be moved beneath the x axis. Jitter functions with both positive max and negative min values are applied for the rest of the charts including the dancability chart, energy chart, liveness chart, and speechiness chart.

The white background is chosen for its neutral nature and suitability for contrasting bright color of data points after testing alternate colors like light gray. There are no grid lines indicating specific x-y values in the background because the pattern demonstrated is qualitative data showing a general trend rather than quantitative data relevant to individual data numbers. However, x-axis and y-axis do contain number values on the scale to help refer to the general trend.

d) The Story:

The team focuses on understanding how key attributes, expressed as percentages, correlate with stream counts and popularity on the platform. By exploring each attribute's impact on streaming performance, the team's visualization aims to uncover key patterns that may influence a song's success. This data-driven approach allows the team to visually analyze which features contribute most significantly to high streaming numbers, providing deeper insights into what makes a song resonate with listeners. For different musical features, the stream count is different as the feature percentage goes up.

Based on the analysis of the scatter plots, it can be concluded that popular songs on Spotify tend to exhibit specific musical characteristics. High-energy tracks with moderate danceability are common among the most-streamed songs, while acousticness is generally low, suggesting that highly produced and electronically enhanced tracks are preferred by mainstream listeners. Instrumentalness and speechiness are also low among the most successful songs. Attributes such as valence or positivity, liveness, and danceability, concentrating on the left end of the x axis, do not show a strong, perceptible direct correlation with stream counts as it neither increases as the streams count increase nor decrease vice versa. Overall, the presence of vocals, high energy, and a lower acoustic component appear to be significant factors for streaming success on Spotify.

It is surprising that valence or positivity and liveness do not significantly contribute to the popularity of music shown in stream count because those attributes are always considered to be stereotypically related to pop musc are generalized to be lively and positive in sound perception. The degree of factor on danceability is also surprising to a lesser degree, because stereotypically popular music genres like R&B and many pop songs are popular in dancing competitions, conventions and video.

The data visualization project shows to the audience that the popularity of a song depends high on energy and low on acoustic and instrumental section, and speechiness. Most popular music, therefore, is supposed to be energetic, moderately danceable but low on acoustic and instrumental sections, and lyrics. For music producers and singers, writing music with energetic vocals, minimal spoken word content, and fewer instrumental segments may help their music become more popular.

e) Contribution:

Yining Wang did the data processing work by locating and importing the data and converting string number data into numbers. Chenyu Yu did the visualization work by using d3 to create the scale and dots corresponding to the data and wrote the milestone 2 report. Yiwei Luo wrote the milestone 1 and final deliverable report and diversified the fill color of dots on each chart as well as added random jitter function based on scale limit to the plot based on the feedback from milestone 2.

Source Citation:

Abdullah, M. (2024, September 7). *Spotify has most streamed songs*. Kaggle. https://www.kaggle.com/datasets/abdulszz/spotify-most-streamed-songs?resource=download.