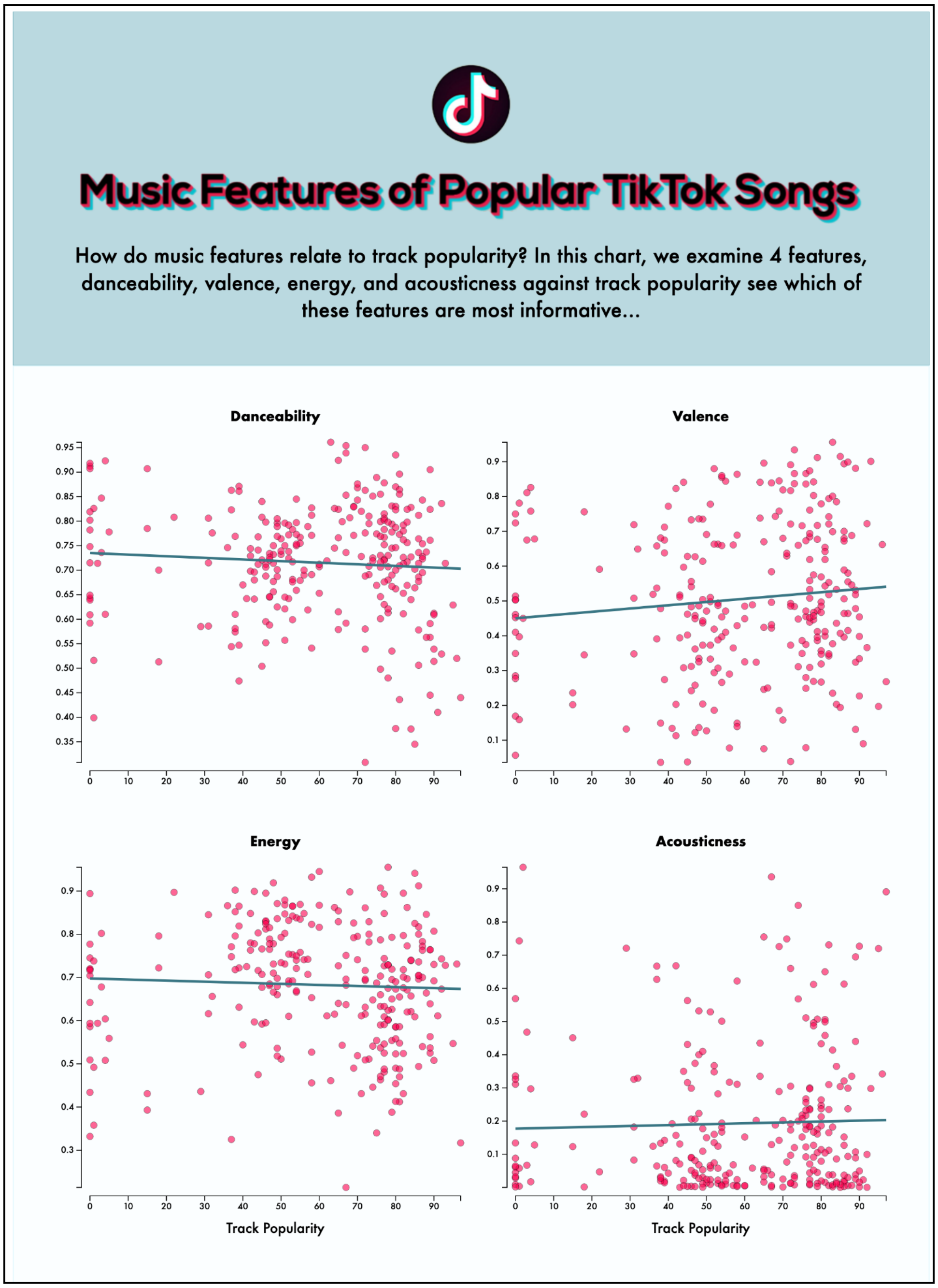
Info 3300: Project 1

Noorejehan Umar, Sydney Bednar, Victoria Eshun

a) At least one screenshot of your final, static visualization.

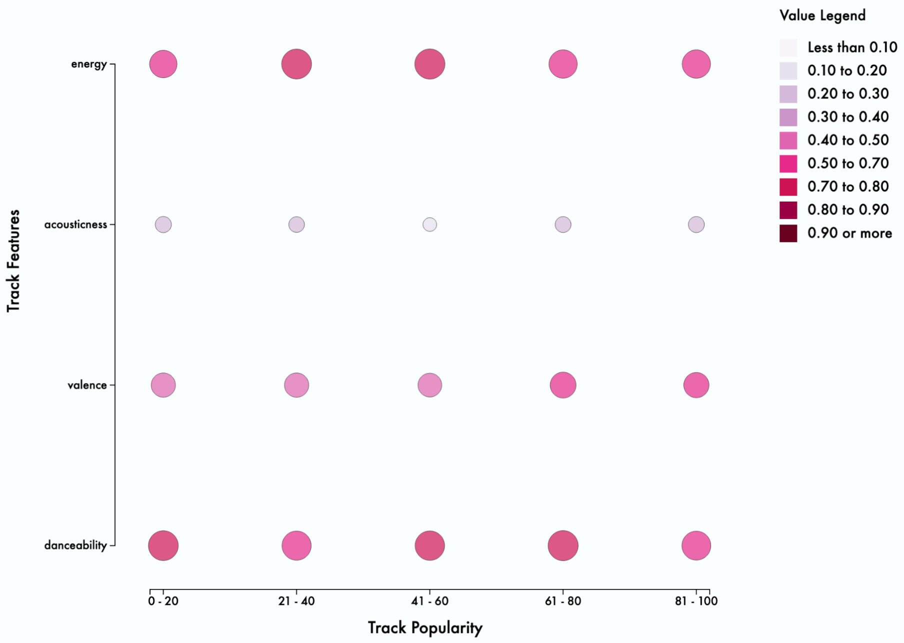


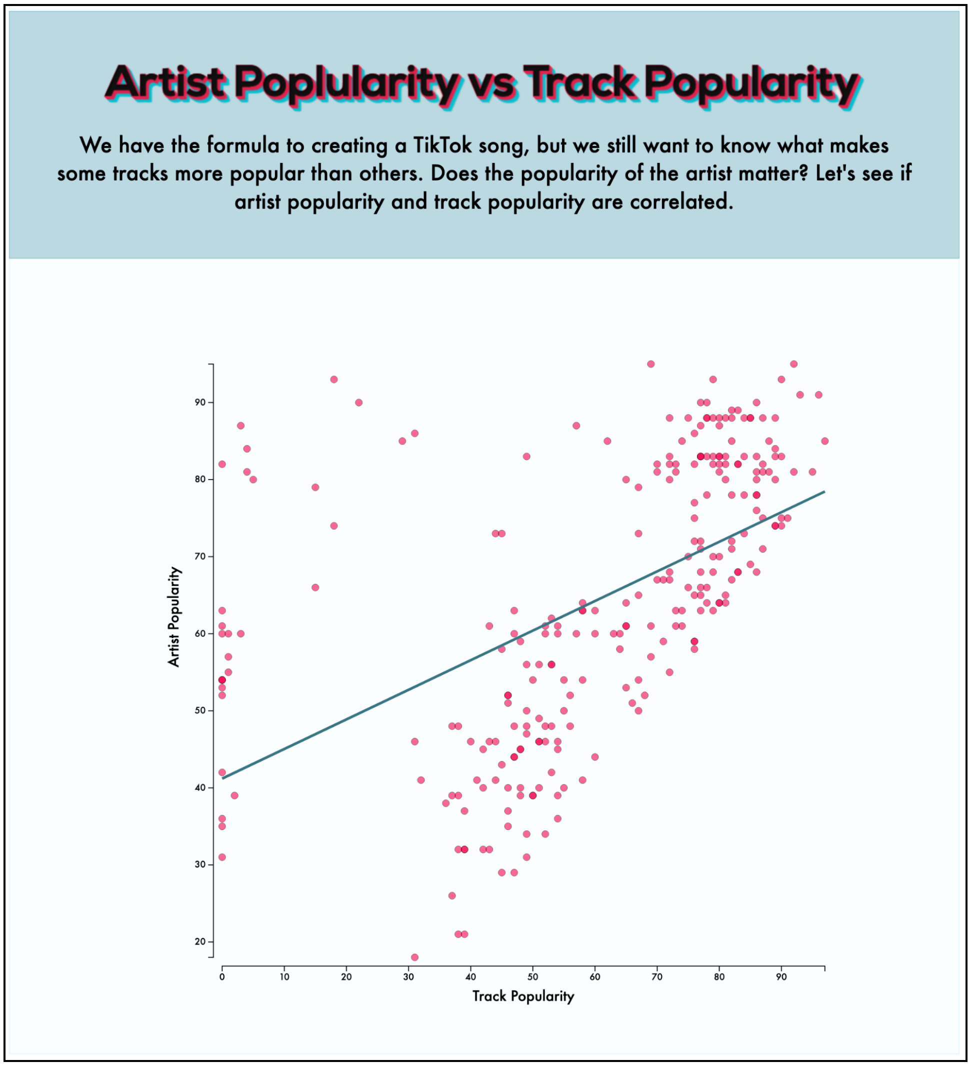
Music Features of Popular TikTok Songs

How do music features relate to track popularity? In this chart, we examine 4 features, danceability, valence, energy, and acousticness against track popularity see which of these features are most informative...

Is there a formula to TikTok songs?

In the charts above, the regression lines have slopes close to zero. Does this mean that all TikTok songs are the same regardless of popularity? Below, we look at the features all together to see what mixture of music features makes a true TikTok song...





Artist Poplularity vs Track Popularity

We have the formula to creating a TikTok song, but we still want to know what makes some tracks more popular than others. Does the popularity of the artist matter? Let's see if artist popularity and track popularity are correlated.

b) A description of the data. Report where you got the data. Describe the variables. If you had to reformat the data or filter it in any way, provide enough details that someone could repeat your results. If you combined multiple datasets, specify how you integrated them. Mention any additional data that you used, such as shape files for maps. Pre-processing is important! You are not required to use every part of the dataset. Selectively choosing a subset can improve usability. Describe any criteria you used for data selection.

The dataset we are using for this project includes entries for 258 popular TikTok songs from 2022. These songs were pulled from the Spotify API by the author who published the dataset here. The Spotify API lets users pull a number of audio features, and we were interested in looking at the relationship between song popularity and these different features. For our project, we have decided to focus on the following columns:

- **artist_pop:** popularity of the artist stated as a value between 0 and 100 as calculated by the Spotify algorithm

- **track_pop:** popularity of the track stated as a value between 0 and 100 as calculated by the Spotify algorithm
- **danceability:** how danceable a track is based on a combination of features including tempo, rhythm stability, beat strength, and overall regularity stated as a value between 0 and 1
- **energy:** how intense and active a track feels based on features including tempo and loudness stated as a value between 0 and 1
- **acousticness:** confidence measure stated as a value between 0 and 1 where 1 is high confidence that the track is acoustic
- **valence:** how positive the music is stated as a value between 0 and 1 where 1 is positive and 0 is negative

We chose to focus on these features because we were thinking about the use case of TikTok to create dancing videos. We hypothesized that danceability, energy, and valence would be close to 1 for many of the most popular songs and that acousticness would be close to 0. Additionally, each of these features use the same scales, and we felt that this would make for a more interesting data visualization when we put the features together. In this visualization, we decided to focus on quantitative data rather than values such as the artist name or track title, which were provided in the original dataset. We also wanted to see whether artist popularity had a significant impact on how popular these songs were compared to each other, or if it was more dependent on the features (danceability, energy, acousticness, valence) of the songs.

c) An overview of your design rationale. A good rule of thumb to follow is “every pixel must be justified.” Instead of a 100,000-element breakdown, give us an overview of the design decisions you made and the trade-offs inherent in how you displayed the data. This part ought to include a description of the mapping from data to visual elements. Describe marks and channels you employ such as position, color, or shape. Mention any transformations you performed, such as log scales.

For our project, we created three different visualizations. Our first visualization included four scatterplots which plotted different features of music against track popularity. We utilized dots as our marks and our channels were their horizontal and vertical positions on an aligned scale. Each dot corresponds to one track or data point. The vertical position of a dot reflects its value for a given feature, while its horizontal position indicates the popularity of the given track. We chose a scatter plot because it's a very well known method of plotting values against each other and this makes the visualization easy to understand and analyze by any one who views it.

For our second visualization, we moved away from a simple scatterplot and took a different approach to visualizing our data. Our aim was to determine if there was a particular combination of features that made songs go viral on TikTok. With our dataset, we created 5 bins to separate the tracks into ranges of popularity using the d3.bin() function. We then computed the average values for each of our four selected features across all the tracks in a given popularity range. We used circles as our marks and size and color as our visual channels. With track features on the y-axis and popularity on the x-axis, each row of circles corresponds to a given track feature while each column corresponds to a popularity range. The size of a circle relative to the other circles in its respective row indicates the average value of that feature for all the tracks in the corresponding popularity range. An increase in circle radius corresponds to an increase in average value for a given track feature. For the circle radii, we used a square root scale which took the average values of the given features and mapped them to a range of 1 to 20. We did this to make the size differences between circles more visible. In addition, we used a sequential color scale with varying saturation and luminosity to complement size as a visual channel. We included a value legend to indicate that values less than 0.10 mapped to a very light purple while values higher than 0.9 mapped to a dark burgundy.

Finally, our third visualization is very similar to the first group of visualizations. We utilized a simple scatter plot with dots as our marks and horizontal and vertical positions on aligned scales as our visual channels. We used linear scales to map track popularity and artist popularity values onto the x and y axes respectively. The vertical position of a dot indicates the popularity of the respective track while the horizontal position indicates the popularity of the artist who made that track.

d) The story. What does your visualization tell us? What was surprising about it? What insights do you want to convey to the viewer of your visualization?

We started by exploring the relationships between track popularity and the various features of popular songs, namely danceability, valence, acousticness and energy, through scatterplots and regression lines. We wanted the viewer to be able to visualize whether any of these features had significant correlation with how popular the song was. Our graphs showed that none of the features seemed to have significant relationships with track popularity. At first, we were surprised to find that none of these features were highly correlated with track popularity. The plots seemed randomly scattered, and the regression lines confirmed that there was no significant linear relationship. We determined that this must be because our dataset was of songs that were similar in nature across these different features already, because they were all songs popular on Tiktok, most of which have high energy, danceability, etc. due to the nature of TikTok being a platform for creating dancing videos.

To confirm our suspicion, we wanted to determine whether there was a ‘formula’ to all TikTok songs, so we looked at all of the features side by side, binned by track popularity on the x-axis with each feature labeled on the y-axis. We used a color scale as well as circle radius to differentiate average values for each feature in a bin. Though there are slight differences in features across bins, overall we see that on average, most songs have the same mixtures of feature levels despite differences in track popularity.

Since we found no significant relationship between track popularity and features, we decided to explore whether artist popularity affected how popular a track was. We made a scatterplot to see what this relationship looked like and found a positive linear relationship. Our results suggest that more popular tracks correlate with more popular artists, regardless of what features those songs have.

Based on all of our results, we came to the conclusion that on Tiktok, popular songs do not necessarily become more popular than others based on any features that the song itself has, such as valence, energy, danceability and acousticness. This could be because popular songs do not differ significantly from each other based on these features. However, the more popular the artist is, the more popular the song will be on Tiktok. This makes sense because more people tend to follow these artists and circulate their music in their own videos.

e) At the end of your PDF file, also include an outline of team contributions to the project. Identify how work was broken down in the group and explain each group member's contributions to the project. Give a rough breakdown of how much time you spent developing and which parts of the project took the most time.

Ideation and Brainstorming (4-5 days):

Sydney: Found 3 different datasets and wrote summaries on what each of them were. Once we voted on our top three, came up with domain tasks and sketches for each visualization. Helped think of what features we want to focus on for our visualizations.

Noorejehan: Found 3 different datasets and wrote summaries on what each of them were. Once we voted on our top three, came up with domain tasks and sketches for each visualization. Helped think of what features we want to focus on for our visualizations.

Victoria: Found 3 different datasets and wrote summaries on what each of them were. Once we voted on our top three, came up with domain tasks and sketches for each visualization. Helped think of what features we want to focus on for our visualizations.

Coding (8-11 hours):

Sydney: Worked on the scatterplots for visualizing the relationship between track popularity and different features. Worked on the main plot for comparing all features with popularity.

Created color legend and linear regressions. Coded styling for charts and the full visualization output.

Noorejehan: Worked on the scatterplots for visualizing the relationship between track popularity, different features and artist popularity. Worked on the main plot for comparing all features with popularity.

Victoria: Worked on the scatterplots for visualizing the relationships between track popularity and the different features. Worked on the main plot for comparing averages of the four features across popularity ranges.

Writing (5-6 hours):

Sydney: Worked on describing our dataset and hypothesis for the writing portion.

Noorejehan: Worked on describing the story behind our visualizations for the writing portion.

Victoria: Worked on describing the design decisions behind our visualizations for the writing portion.