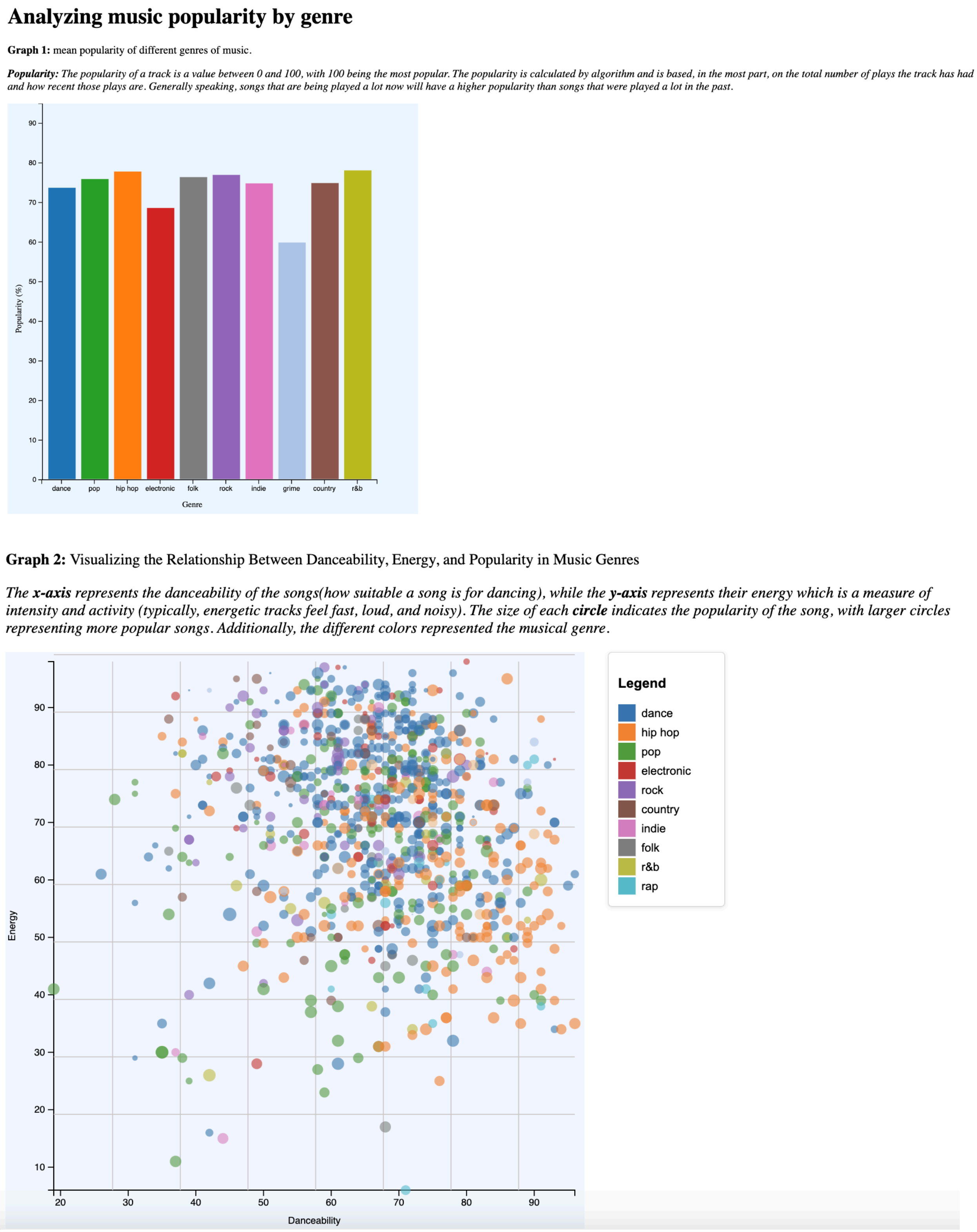
Final Report

Screenshots of the final visualization



Analyzing music popularity by genre

Graph 1: mean popularity of different genres of music.

Popularity: *The popularity of a track is a value between 0 and 100, with 100 being the most popular. The popularity is calculated by algorithm and is based, in the most part, on the total number of plays the track has had and how recent those plays are. Generally speaking, songs that are being played a lot now will have a higher popularity than songs that were played a lot in the past.*

Graph 2: Visualizing the Relationship Between Danceability, Energy, and Popularity in Music Genres

*The **x-axis** represents the danceability of the songs(how suitable a song is for dancing), while the **y-axis** represents their energy which is a measure of intensity and activity (typically, energetic tracks feel fast, loud, and noisy). The size of each **circle** indicates the popularity of the song, with larger circles representing more popular songs. Additionally, the different colors represented the musical genre.*

Description of data

We looked at several datasets in search of the “perfect” one. We ended up choosing a dataset from Kaggle based on Spotify created by Michael Morris.
The dataset includes the following variables:
['title', 'artist', 'top genre', 'year released', 'added', 'bpm', 'nrgy', 'dnce', 'dB', 'live', 'val', 'dur', 'acous', 'spch', 'pop', 'top year', 'artist type']
Using libraries available to Python, we dropped 3 rows that strictly contained NaN values. We modified ‘top genre’ significantly. The original dataset (spotify-2010-2019.csv) included genre types that were often too niche or granular. For example, instead of falling under “pop” (genre) broadly, there were subgenres of pop like “israeli pop”, “colombian pop”, and so on.
In total, there were 132 genres. After mapping and merging as many genres as possible, we reduced that number to 39.
Besides minimizing potential overload of information that would make graphs focusing on genre difficult to interpret, this also helps deal with potential data skewing where niche genres would have an overinflated presence in a graph. Of course, this comes with the trade-off such as loss of data granularity, but for our purposes, the trade-off is outweighed by the quality of the final product.
Similarly, we removed rows where the count of songs for a given genre was less than 10. This is because there were multiple genres that included just one song.
We also renamed “top genre” to “top-genre” just so we would not have any issues because of the space.
The modified dataset is called filtered.csv.
The filtering process is documented in a Jupyter file called data_filtering.ipynb included in the submission.

Rationale

Visualization 1

In this data visualization, we aimed to represent the mean popularity of different music genres by creating a bar chart. Each bar’s height meant to represent the average popularity of tracks within a specific genre, making it easy to compare how different genres rank in terms of popularity

Mapping

The x-axis represents different music genres, and the data is mapped using scaleBand() which will assign each genre an equal amount of space on the axis. This added to the readability of chart by making all bars uniform in size and spacing.

Popularity is mapped on the y-axis, it is scaled linearly. The linear scale was chosen over alternatives like a logarithmic scale because the values were already in a narrow range (0-100), a log scale wouldn’t be suitable for this as it might obscure the differences in popularity.

Marks and Channels

Marks: The bars are the primary visual marks used to represent each genre's popularity. These rectangular bars provide a clear way to compare relative popularity

Channels:

- Position is used along the x-axis to represent each genre and along the y-axis to show the mean popularity
- Height of the bar was directly proportional to the mean popularity value
- We chose to color-code the bars so that Visualization 2 would have genres of the same colors as Visualization 1. This was done to remain consistent and to make comparison between the two visualizations easier.
- Width was fixed at 40 units, to ensure uniform bars and avoid too much empty space or overlap

Other Design Additions

Background-color: We chose a light blue background color in order to draw attention to the graphs. Highlighting each graph in a light blue color added uniformity to the overall piece and increased the overall readability of the file.

Graph title and caption: We chose to add a title and small caption above the graph to further explain the graph if there was any confusion. Some of the fields may be misinterpreted so we decided to add a small description of the axes.

Trade-off

1. Displaying mean popularity: By averaging the popularity values for each genre, the visualization sacrifices insight into the distribution or variance of popularity between the genres. However, we decided this trade-off was necessary to simplify the graph and focus on the general trend of popularity
2. Limiting the number of genres: We chose to display only 10 genres to maintain readability, as including more genres could make the chart more cluttered and harder to interpret. There were 5 more genres however the amount of data for those genres were so small we thought that would make our data less accurate. Although there were subcategories to some large genres such as pop, we were only able to include an umbrella genre instead of each individual sub-genre.

Visualization 2

In this data visualization, we made several design decisions to clearly display the relationship between danceability, energy, and popularity of songs for different music genres.

Mapping

Danceability is mapped the x-axis and Energy is mapped to the y-axis, both are scaled linearly. This decision prioritizes a straightforward interpretation of the data, allowing users to easily compare the relationship between these two features. Typically a logarithmic scale would be used to display data with an expansive value disparity/distribution, which was not the case for this dataset.

The circles represented the individual data points, where each circle corresponded to a specific genre of music and its values for danceability and energy.

Popularity was visualized through the size of the circles, where larger circles indicated more popular songs. I chose to make the circle sized range from 0.5 to 7 since circles larger than that would affect the interpretability of the code and overlap over the other data points. I chose to use a linear scale, which ensured a proportional and clear visual distinction between popular and less popular songs.

Marks and Channels

Marks: Circles are used as marks, their positions were determined by the values for danceability (x-axis) and energy (y-axis). This choice ensures an uncluttered display for the scatterplot, allowing for easy comparison between points

Channels:

- Position is used for mapping the danceability and energy
- Color is employed to differentiate between genres, with each genre assigned to a distinct color using a categorical color scale. This was chosen to enhance the viewer's ability to distinguish genres visually
- The size of the circles represented the popularity of the song, with larger circles indicating higher popularity.

Other Design Additions

Gridlines: Horizontal and vertical gridlines were added to help guide the viewer's eye across the plot, providing reference points for the comparing values. However, to keep the focus on the points, the gridlines were rendered in a subtle, light gray color.

Legend: A legend was placed on the right side of the chart to clarify which colors correspond to which genres. This ensures that the color scheme remains interpretable, even for users unfamiliar with the specific data. We gave it a border so that it would stand out against the background and be more identifiable as a legend.

Background-color: We chose a light blue background color in order to draw attention to the graphs. Highlighting each graph in a light blue color added uniformity to the overall piece and increased the overall readability of the file.

Circle Opacity: We chose to make the opacity 0.6 to allow for overlapping circles to remain visible. Without this transparency, user might not notice clusters or overlapping data points, which would hinder the ability to interpret dense areas of the plot

Graph title and caption: We chose to add a title and small caption above the graph to further explain the graph if there was any confusion. Some of the fields may be misinterpreted so we decided to add a small description of the axes and the importance of circle sizes. We thought it would be intuitive to address what the circle sizes correlate to within the caption/description.

Trade-off

1. Using circle sizes for popularity makes it intuitive to grasp popularity differences, but it can be hard to compare circles accurately when they overlap or when small differences in sizes are subtle
2. Colors were used to differentiate genres, and although the color palette was distinct, using many colors might be overwhelming when there are multiple categories

The Story

The first graph shows which music genres are most popular, and how they compare with other genres. As we transition to the second graph, we are investigating what components of the music make it so popular. We wanted to see if certain popular genres attributed their popularity to the songs' makeup of different components. We chose danceability and energy to see if there was a correlation between these two components of the music and the popularity of songs broken down by genre. Although there isn't a clear winner in terms of most popular genre in the second graph, the bubble size shows the popularity. A surprising finding is that genres like dance and hip-hop did not have a significantly higher energy and danceability rating than other genres as listeners might have assumed. Most genres have songs that range in levels of energy and danceability instead of all songs of one genre being grouped together. We can see from the graph though that the most densely populated area of the graph is from 60-80 in danceability and energy. This seems to be the make up of most songs no matter the genre. This could be due to changes in music trends by representing what properties in music are currently popular. The insights we want to convey to our viewers is that these visualizations show that music doesn't vary that much genre to genre. Although music sounds very different to our ears, the makeup components of vastly different songs are surprisingly similar.

Contribution

How was work broken down

We generally worked together for the project. We broke down certain roles in in-person and zoom meetings. Everyone had different roles according to their individual skill sets. When we had issues with the code we would discuss in a group how we can tackle the issue. Overall, we all contributed equally and communicated effectively. We all juggled different roles like – project manager, designer, and developer. We delegated tasks and followed a project timeline for when everything will be completed [ex. Who took on which visualization, how we would all help and brainstorm if a certain idea wasn't as visually effective for the visualizations].

Each member's contributions

Maggie

- Brainstormed ideas for the final visualizations
- Developed the second visualization
- Wrote the rationale portion of the final report

Yama

- Investigated and explored several datasets. Preprocessed and cleaned the final dataset.
- Wrote the Data Description portion
- Helped with troubleshooting the second visualization.
- Modified the colors of the first Visualization

Kelly

- Developed first visualization
- Demoed our charts in class and collected feedback

- Wrote story portion of the final report
- Added to graph 1 section of final report

How much time was spent developing
It took roughly 6 hours to develop the 2 visualizations. This included troubleshooting, data filtering, and developing the individual visualizations.

Which part of the project took the most time
Developing the visualizations and brainstorming ideas for the visualizations took the majority of the time. We spent two meetings on this, and took a week to decide on final designs.