DSCI Group Project - Milestone I

Projet Name: Spotify Data Visualization

Group Members: Emily Ahn, Weihao (Beren) Sun

Group Name: Beren & Emily’s Spotify Blend

EDA:

See the Jupyter Notebook

Project Scope:

Introduction:

The topic of this project is to visualize the Spotify soundtrack dataset to show the relationship between key features of the songs. Our motivation is to give our visualization readers an intuitive understanding of which soundtrack is more popular, and what are the features like. The features include loudness, popularity, danceability, mode, etc.

Data Analysis:

We use the Spotify Soundtrack dataset for our project, and there are all the variables in the dataset, including all the information.

- track_id: unique ID of each soundtrack. Type: unique key. 89741 values
- track_name: Soundtrack name. Type: nominal. 89741 values
- album_name: Album name. Type: nominal. 89740 values.
- artist_1: The first artist. Type: nominal. 89740 values.
- artist_2: The second artist. Type: nominal. 22584 objects.
- artist_3: The third artist. Type: nominal. 6784 objects.
- artist_4: The fourth artist. Type: nominal. 2124 object.
- track_genre_1: The genre in which the track belongs. Type nominal. 89741 values.
- track_genre_2: The second genre in which the track belongs. Type: nominal. 16641 objects.
- track_genre_3: The third genre in which the track belongs. Type: nominal. 4929 objects.

- track_genre_4: The fourth genre in which the track belongs. Type nominal. 1945 objects.
- track_genre_5: The fifth genre in which the track belongs. Type: nominal. 573 objects.
- Explicit: Whether or not the track has explicit lyrics. Type: nominal. Two levels.
- Mode: The modality (major or minor) of a track. Type: nominal. Two levels.
- Key: The key the track is played in. Type: nominal. 12 levels.
- Time_signature: The (estimated) time signature of the track. Type: nominal. 5 levels.
- Tempo: The overall estimated tempo of a track in beats per minute (BPM). Type: quantitative. Range: (0,243.37)
- Duartion_s: The track length in seconds. Type: Quantitative. Range: (0,5237.30)
- Popularity: The popularity of a track. Type: Quantitative. Range: (0,100)
- danceability: Danceability describes how suitable a track is for dancing based on a combination of musical elements including tempo, rhythm stability, beat strength, and overall regularity. Type: Quantitative. Range: (0,1).
- Loudness: The overall loudness of a track in decibels (dB). Type: quantitative. Range: (-49.531,4.532).
- Speechiness: Speechiness detects the presence of spoken words in a track. Type: quantitative. Range: (0,1).
- Acousticness: A confidence measure from 0.0 to 1.0 of whether the track is acoustic. Type: quantitative. Range: (0,1).
- Instrumentalness: Predicts whether a track contains no vocals. Type: quantitative. Range: (0,1)
- Liveness: Detects the presence of an audience in the recording. Type: quantitative. Range: (0,1)
- Energy: a perceptual measure of intensity and activity. Type: quantitative. Range: (0,1)
- Valence: describes the musical positiveness conveyed by a track. Type: quantitative. Range: (0,1)
-

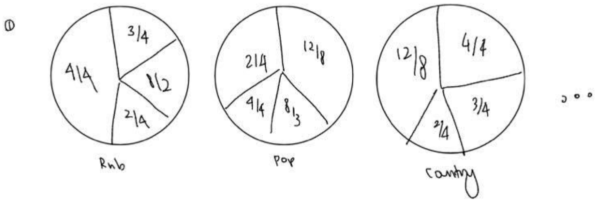
Task Analysis:
Audience: song writers
questions:

- Which key is the most popular in each genre? (Find extreme)
- Rank the top 50 most popular songs. (Sort)
- Is there a correlation between the audio features and popularity? (Correlate)
- What is the range of song lengths by genre? (determine range)
    -

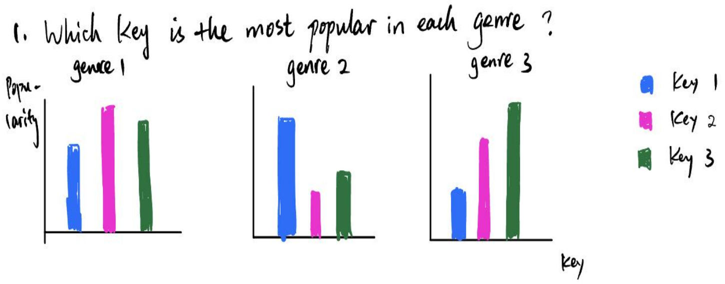
- Rank the most popular genre (Sort)
    -
- What is the tempo distribution of each mode? (Characterize Distribution)

Preliminary Sketches:

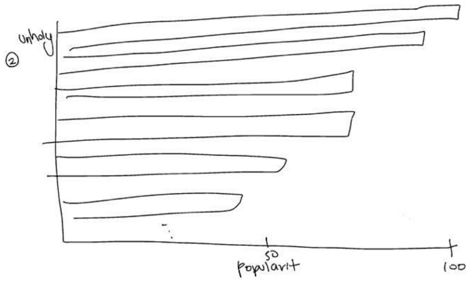
- **Which key is the most popular in each genre? (Find extreme)**
- Emily



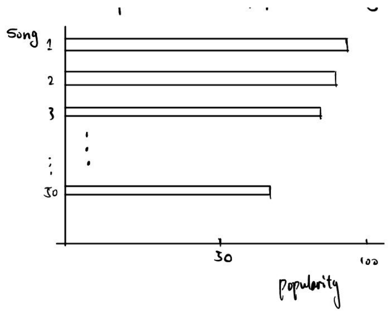
Weihao



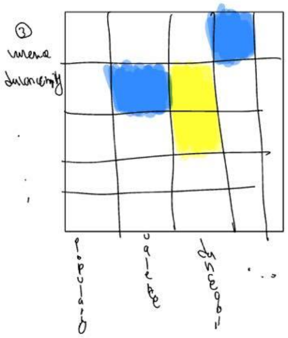
- **Rank the top 50 most popular songs. (Sort)**
- Emily



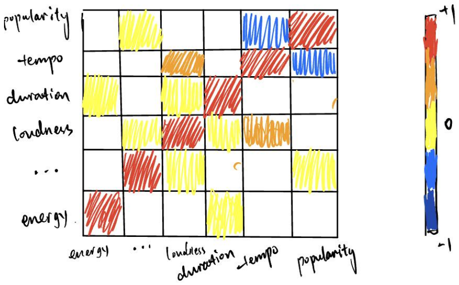
Weihao



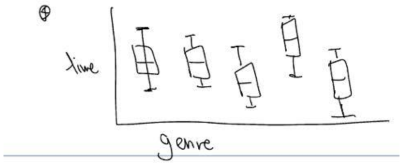
- **Is there a correlation between the audio features and popularity? (Correlate)**
- Emily



- Weihao

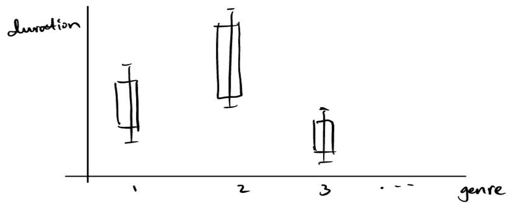


- **What is the range of song lengths by genre? (determine range)**
- Emily

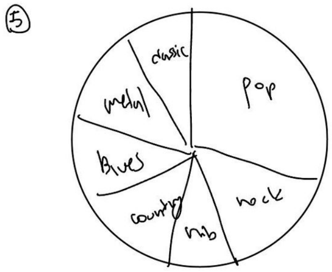


-

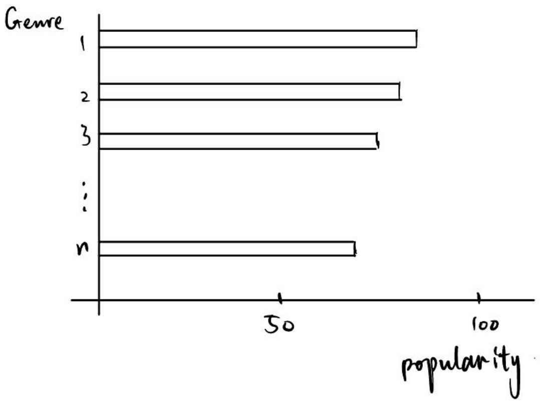
Weihao



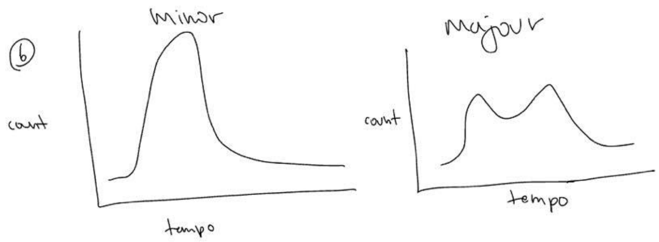
- **Rank the most popular genre (Sort)**
- Emily



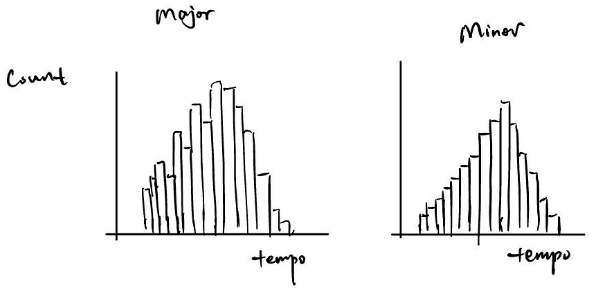
- Weihao



- **What is the tempo distribution of each mode? (Characterize Distribution)**
- Emily

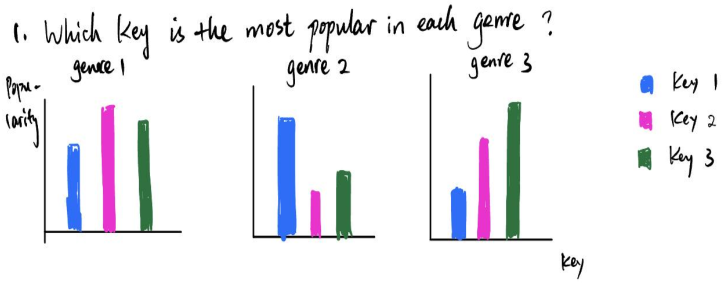
- 

Weihao



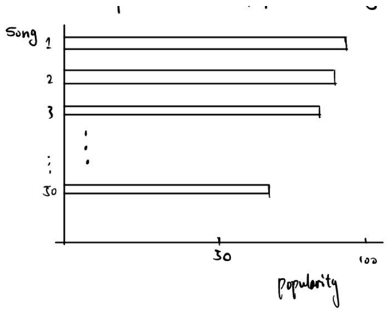
Selected Sketches:

- **Which key is the most popular in each genre? (Find extreme)**
  Selected sketch:

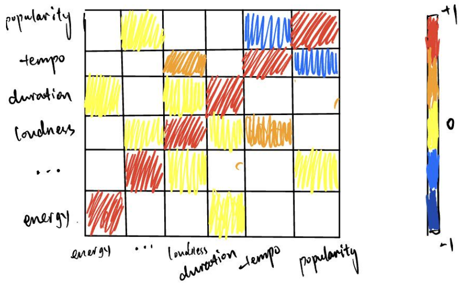
  We select this sketch because this visualization can address the question, which is to find the most popular key in each genre. Using a bar chart is clear and easy to compare the popularity among keys. Also, using colours to represent different keys gives a more intuitive way to recognize different keys in this visualization. To represent the key popularity in a different genre, we use facets to create one plot for each genre, which avoids concentrating all the bars into one large plot. One theory that this visualization can apply is to use multiple views. The partition of the plots is side-by-side, which gives a clear view for the readers. In the final prototype, we will add some tooltips and other interaction types into the visualization to provide some selection or hover functions.

- **Rank the top 50 most popular songs. (Sort)**
  Selected sketch:



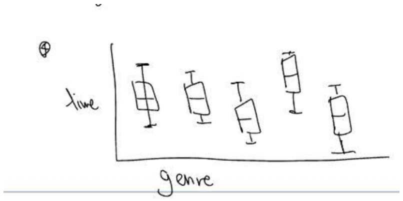
We choose this visualization because using a bar chart is suitable for visualizing the rank of the song's popularity. It is clear to read the bar plot from the top bar to the bottom, which represents the song with the largest popularity and the song with the least popular among the top 50 ones. It is also easy to compare the popularity of songs by looking at the bar length from the top to bottom. This visualization perfectly addresses one lower-level task by using Stasko's taxonomy, which is the sorting problem as we already stated.

- **Is there a correlation between the audio features and popularity?**
  Selected sketch:

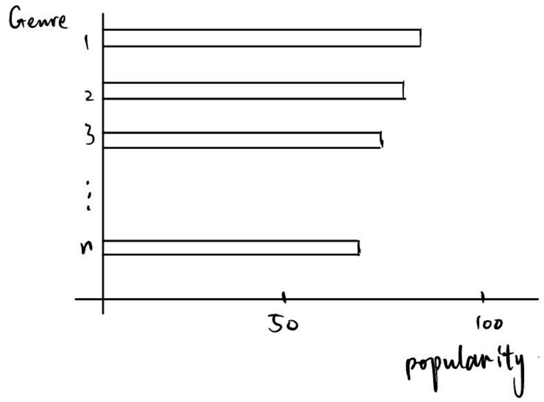


We choose this sketch to address the problem of finding correlations between audio features and popularity. This is because all of the features as well as the popularity are quantitative variables. Therefore, a correlation matrix is the most suitable visualization for such a task. In this visualization, all the variables are on the X and Y axis, and each grid represents the correlation between the two specific variables. Colours represent the strength of correlations, which makes the matrix more intuitive. In the final prototype, we will add the correlation values as well as tooltips to help the user to get the exact correlation level. We will also change the colour to a more meaningful colour set to represent such an ordinal number, such as using increasing luminance as stated in class.

- **What is the range of song lengths by genre? (determine range)**

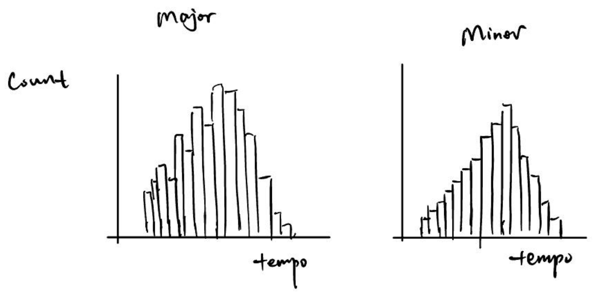


- The box plot is suitable since it will show a graphical representation of the distribution of a dataset through the box's lines that show the lower range and the upper range as well as the median value. There will be genre on the x-axis and time by minutes and seconds on the y-axis. Aside from the interquartile range viewers will be able to distinguish the minimum and maximum song lengths and compare between genres and its range in length of songs. It will also plot some of the outliers to show observations that are significantly different from the rest of the data. Another idea we had while researching box plots was a violin plot which shows the density and distribution in the data hence might show more details than just a regular boxplot, so it may be a consideration once we start plotting.
- **Rank the most popular genre (Sort)**



- We chose this visualization because although the pie chart shows relative contribution to the total popularity by each genre, but a bar graph makes more sense since it will be able to show the mean popularity of each genre as well as be able to compare relative to other genres. Each genre will be placed on the y-axis and sorted by the mean popularity across each genre. It will be easy to read with clear labelling which makes it an ideal choice for users who are trying to figure out the most popular genre.

- **What is the tempo distribution of each mode? (Characterize Distribution)**



- We chose this visualization of a histogram where we are showing the frequency of tempo within a given bin for each mode. Having the two histograms allows viewers to be able to compare the distribution between the two modes. By having a histogram it'll allow viewers to identify patterns and trends easily to make informed decisions. Having two histograms side by side avoids bias and shows if there are correlation between mode and tempo since music in minor mode tends to be sadder so may lead to a lower tempo in the music. It also does not contain any unnecessary designs and is simple with multiple views. In this graph it will be important to choose the right bin value so that it will show the pattern of the distribution well without being too busy.