

Milestone I - Project Scope

Group: Full-Mark Musician

Group Member: Linda Huang, Jialu Jin, Minghao Wang

Project Name: Spotify Data Exploring

Introduction

The dataset we use contains over 125 different genres of Spotify tracks. The dataset contains attributes like track name, duration, and song energy/loudness that can be used to build a recommender system or used for classification purposes.

For our project, we are interested in exploring the relationship between different attributes of Spotify songs. While big data is great for information accuracy, we will only be focusing on exploring 10 randomly selected genres.

Data Analysis

The dataset is a structured dataset as it is stored in a tabular format with 89741 rows and 27 columns. Among 27 attributes, there are 15 categorical attributes and 12 quantitative attributes. The dataset contains the following attributes or features:

- track_id: a string that represents the Spotify track ID which is the unique identifier of each track
- track_name: a string that represents the name of the track
- album_name: a string that represents the album name in which the track appears
- artist_1: a string that represents the name of the artist who performed the track
- artist_2: a string that represents the name of the second artist who performed the track
- artist_3: a string that represents the name of the third artist who performed the track
- artist_4: a string that represents the name of the fourth artist who performed the track
- track_genre_1: a string that represents the genre in which the track belongs
- track_genre_2: a string that represents the second genre in which the track belongs

- track_genre_3: a string that represents the third genre in which the track belongs
- track_genre_4: a string that represents the fourth genre in which the track belongs
- track_genre_5: a string that represents the fifth genre in which the track belongs
- explicit: a boolean that represents whether or not the track has explicit lyrics (TRUE = yes it does; FALSE = no it does not OR unknown)
- mode: a string that represents the modality (major or minor) of a track, i.e. the type of scale from which its melodic content is derived
- key: a string that represents the key the track is played in. E.g. C, C-sharp, D, E-flat, E, F, and so on. If no key was detected then the value is NA
- time_signature: an int that represents the (estimated) time signature of the track. The time signature (meter) is a notational convention to specify how many beats are in each measure (or bar). The value is an integer that ranges from 3 to 7 indicating time signatures of 3/4 to 7/4
- tempo: a float that represents the overall estimated tempo of a track in beats per minute (BPM). The tempo is the speed or pace of a given track and derives directly from the average beats per minute
- duration_s: a float that represents the track length in seconds
- popularity: an int that represents the popularity of a track. This is a value between 0 and 100, with 100 being the most popular. The popularity is calculated by an algorithm from Spotify
- danceability: a float that describes how suitable a track is for dancing based on a combination of musical elements including tempo, rhythm stability, beat strength, and overall regularity. A value of 0.0 is least danceable and 1.0 is most danceable
- loudness: a float that represents the overall loudness of a track in decibels (dB)
- speechiness: a float that detects the presence of spoken words in a track. The more exclusively speech-like the recording (e.g. talk show, audio book, poetry), the closer to 1.0 the attribute value. Values above 0.66 describe tracks that are probably made entirely of spoken words. Values between 0.33 and 0.66 describe tracks that may contain both music and speech, either in sections or layered, including such cases as rap music. Values below 0.33 most likely represent music and other non-speech-like tracks
- acousticness: a float that represents a confidence measure from 0.0 to 1.0 of whether the track is acoustic. 1.0 represents high confidence the track is acoustic
- instrumentalness: a float that predicts whether a track contains no vocals. "Ooh" and "aah" sounds are treated as instrumental in this context. Rap or spoken word tracks are clearly "vocal". The closer the instrumentalness value is to 1.0, the greater likelihood the track contains no vocal content

- liveness: a float that detects the presence of an audience in the recording. Higher liveness values represent an increased probability that the track was performed live. A value above 0.8 provides strong likelihood that the track is live
- energy: a float that measures from 0.0 to 1.0 and represents a perceptual measure of intensity and activity. Typically, energetic tracks feel fast, loud, and noisy. For example, death metal has high energy, while a Bach prelude scores low on the scale
- valence: a float that measures from 0.0 to 1.0 describing the musical positiveness conveyed by a track. Tracks with high valence sound more positive (e.g. happy, cheerful, euphoric), while tracks with low valence sound more negative (e.g. sad, depressed, angry)

Categorical:

- track_id
- track_name
- album_name
- artist_1
- artist_2
- artist_3
- artist_4
- track_genre_1
- track_genre_2
- track_genre_3
- track_genre_4
- track_genre_5
- explicit
- mode
- key

Number of levels for categorical data:

- track_id: 89741
- track_name: 73609
- album_name: 46590
- artist_1: 17649
- artist_2: 11217
- artist_3: 4110
- artist_4: 1447
- track_genre_1: 113
- track_genre_2: 113
- track_genre_3: 88

- track_genre_4: 58
- track_genre_5: 33
- explicit: 2
- mode: 2
- key: 12

Quantitative:

- time_signature
- tempo
- duration_s
- popularity
- danceability
- loudness
- speechiness
- acousticness
- liveness
- energy
- valence

Range for quantitative data:

- time_signature: 0 to 5
- tempo: 0 to 219.97
- duration_s: 8.59 to 4447.52
- popularity: 0 to 83
- danceability: 0 to 0.981
- loudness: -38.409 to 3.15
- speechiness: 0 to 0.924
- acousticness: 0 to 0.996
- instrumentalness: 0 to 0.999
- liveness: 0.0166 to 0.989
- energy: 0.00002 to 0.999
- valence: 0 to 0.994

Columns that contain missing values:

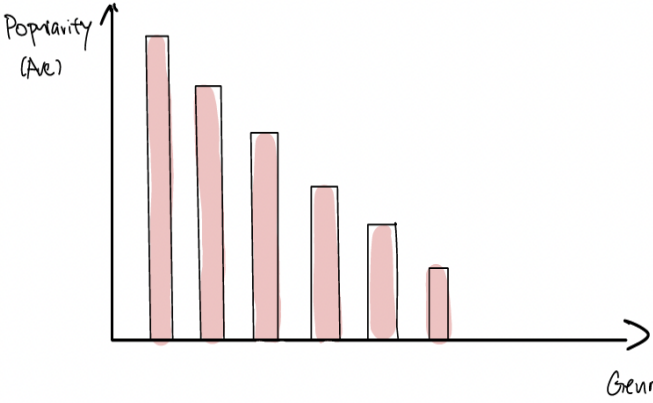
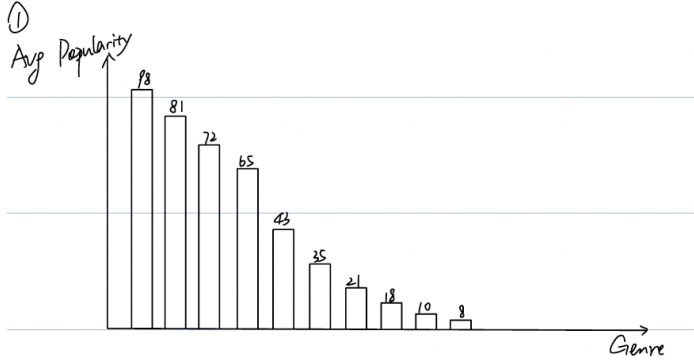
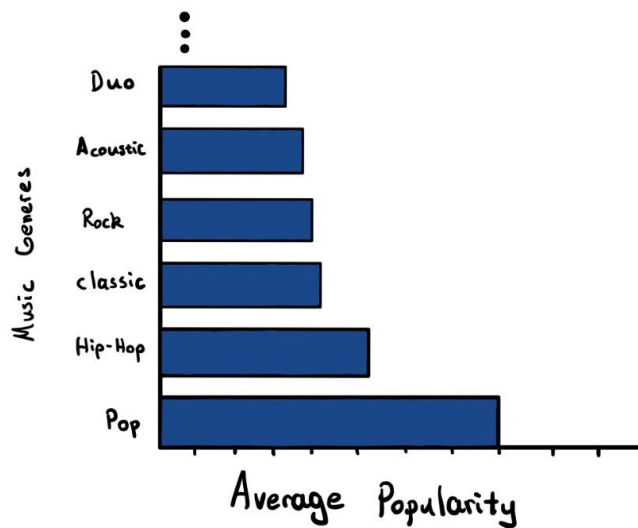
'artist_2','artist_3','artist_4','track_genre_2','track_genre_3','track_genre_4','track_genre_5'

Columns that does contain missing values:

'track_genre_1','explicit','mode','key','time_signature','tempo','duration_s','popularity','danceability','loudness','speechiness','acousticness','instrumentalness','liveness','energy','valence'

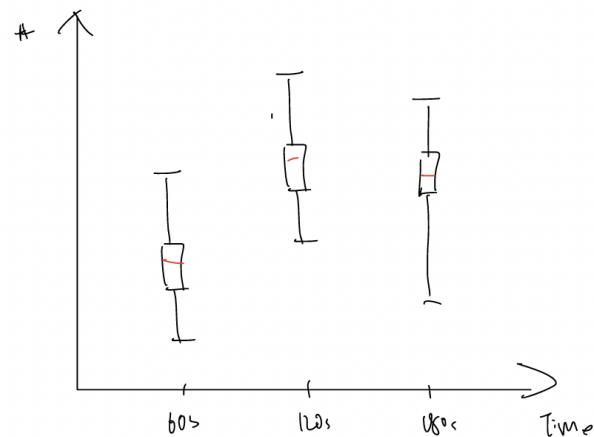
Task Analysis

1. Which **genre** of the song has the highest average **popularity**?
2. For the duration, within 1 mins, 1 - 2 mins, or 2 - 3 mins. Which has the highest median danceability score?
3. What is the difference in the distribution of **popularity** between major and minor **modes**?
4. Is there a relationship between **danceability** and **loudness in different genres**?
5. Is there a difference in the average **tempo** among each **key**?
6. What's the difference between the average **valence** # among different **keys**?

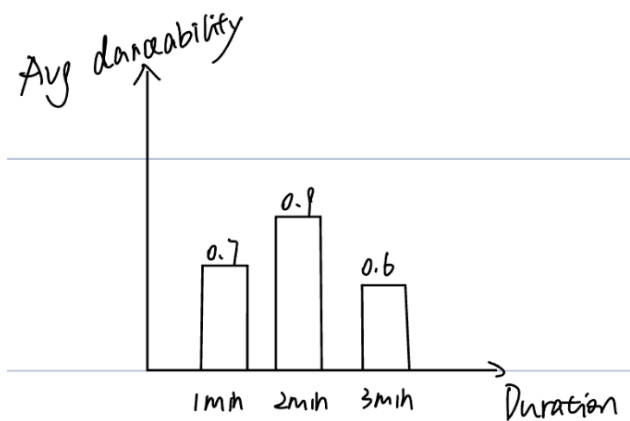
Task Abstraction:	Sketch & Critique :
<p>Which genre of the song has the highest average popularity?</p>	<p>Linda:</p>  <p>Cindy:</p>  <p>Minghao:</p> <p>Which genre of the song has the highest average popularity?</p> 

For duration, with in 1 mins, 1 - 2 mins, or 2 - 3 mins. Which has the highest median danceability score?

Linda:

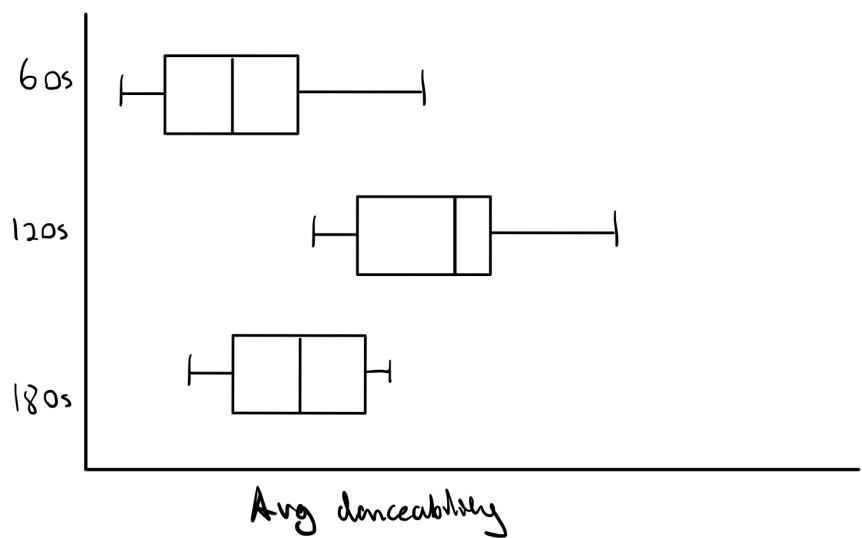


Cindy:



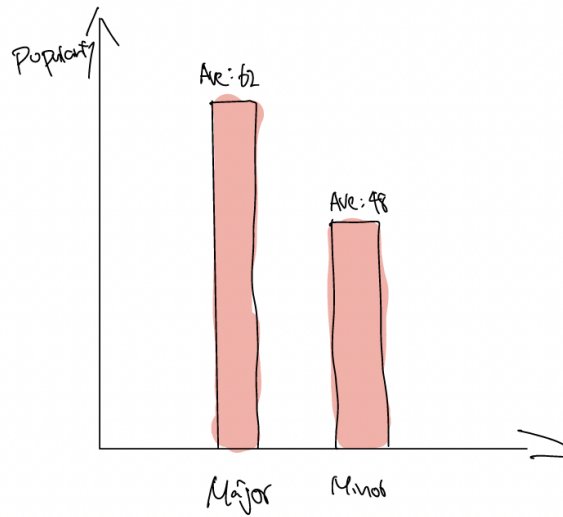
Minghao:

For duration, with in 1 mins, 1 - 2 mins, or 2 - 3 mins. Which has the highest average danceability score?

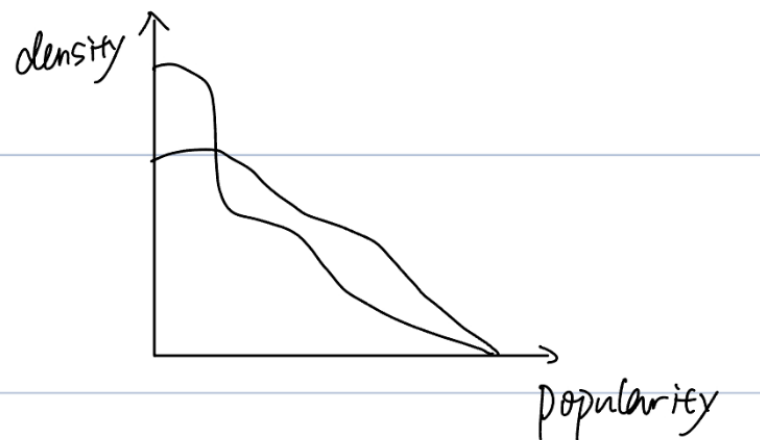


What is the difference in distribution of popularity between major and minor mode?

Linda:

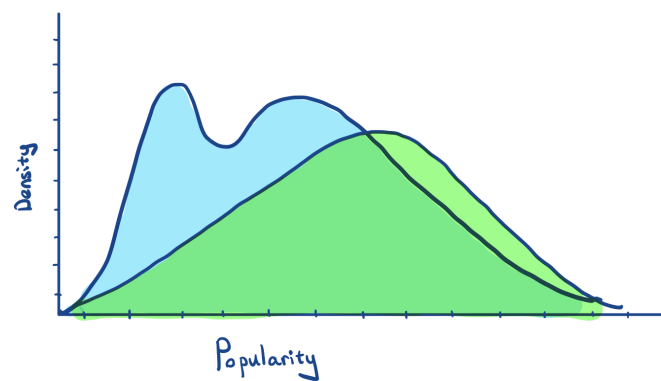


Cindy:



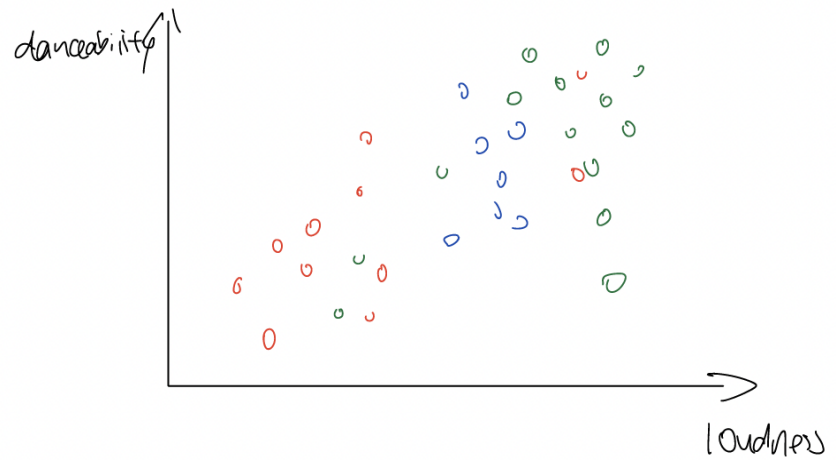
Minghao:

What is the difference in distribution of popularity between major and minor mode?



Is there a relationship between danceability and loudness with different genres?

Linda:

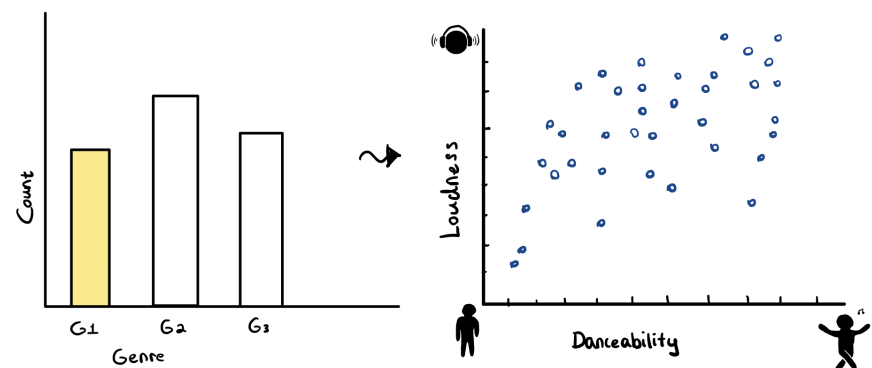


Cindy:



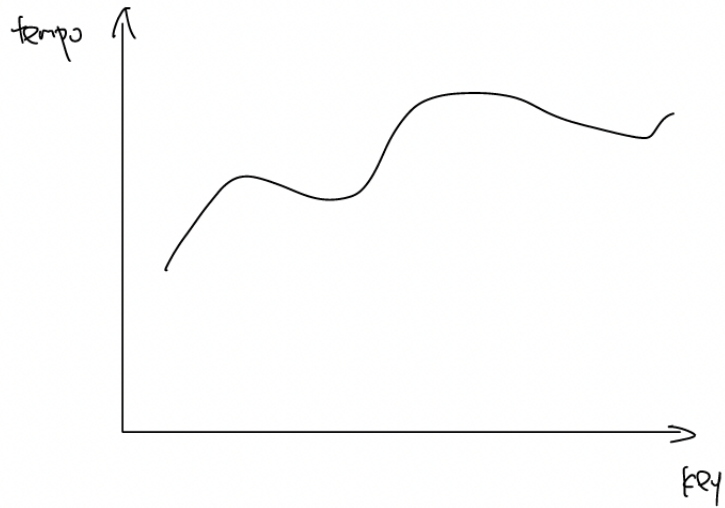
Minghao:

Is there a relationship between dance ability and loudness with different genres?

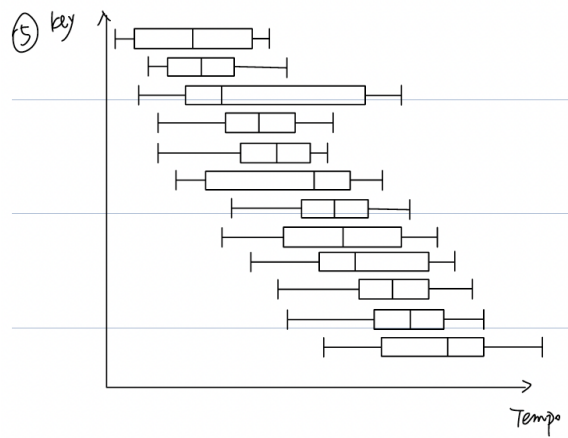


Is there a difference in the average tempo among each key?

Linda:

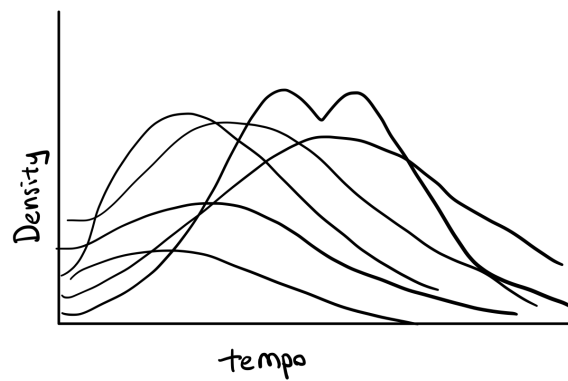


Cindy:



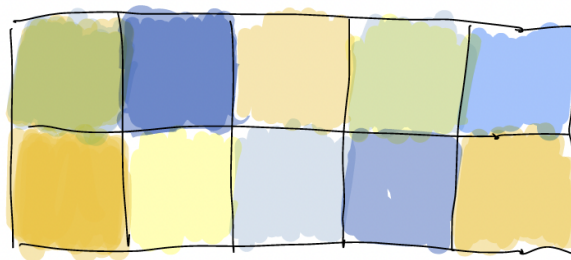
Minghao:

Is there a difference in the average temp among each key?



What's the difference between the average valence # among different keys?

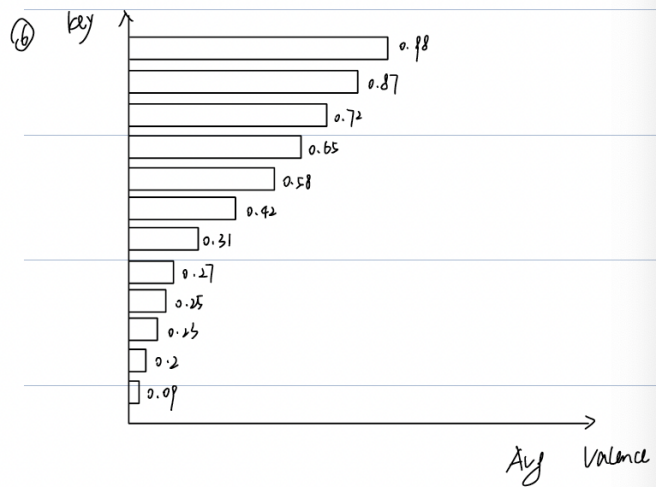
Linda:



Valence:

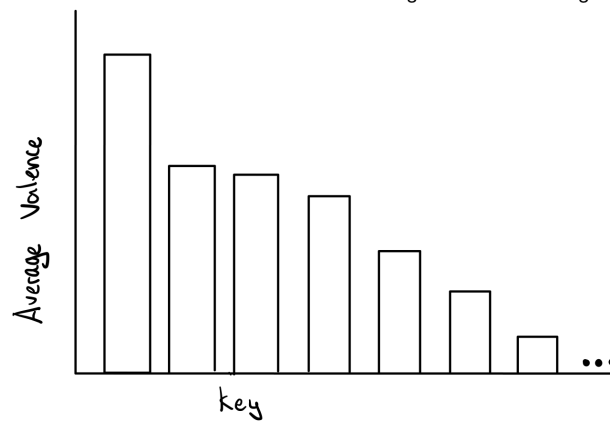


Cindy:



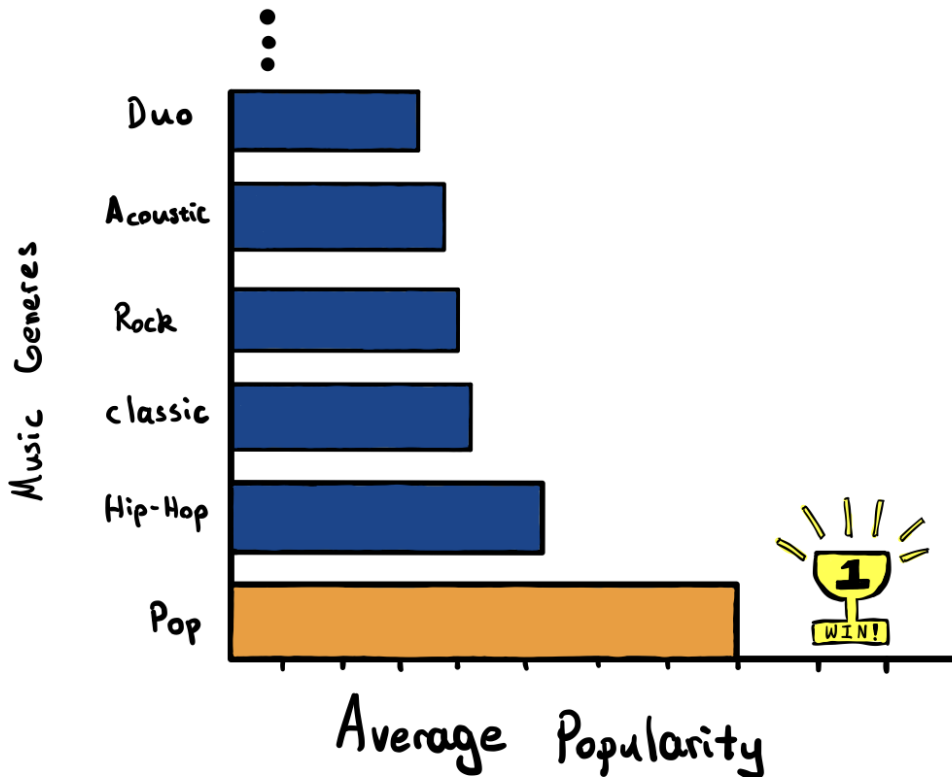
Minghao:

What's the difference between the average valence # among different keys?



Preliminary Sketches

Which genre of the song has the highest average popularity?



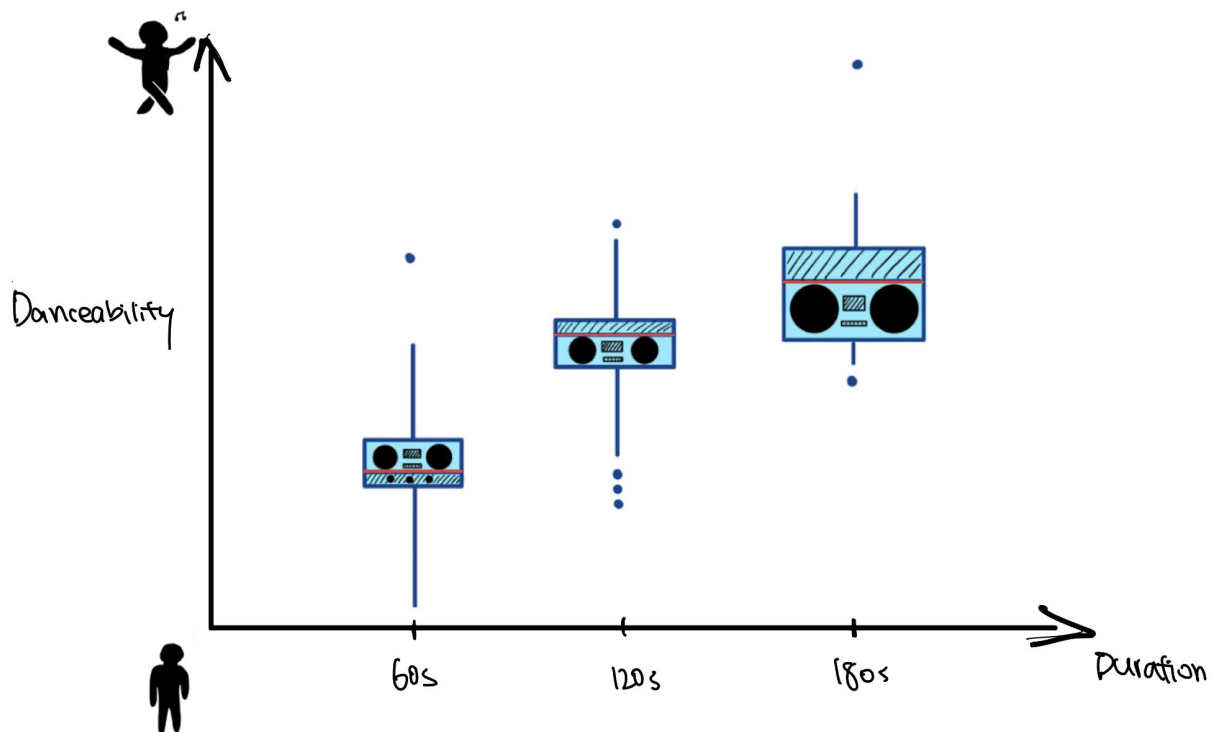
We decided to choose a horizontal bar chart due to its practicality when dealing with lengthy category labels and its ability to enhance readability when comparing the relative sizes of bars. A horizontal bar chart is a highly effective visualization choice when determining which music genre has the highest average popularity. It provides a straightforward and efficient way to compare various categories. The chart's intuitiveness and ease of interpretation make it an excellent tool for conveying data to a broad audience.

The chart has a y-axis that represents the various music genres, with the x-axis representing their respective average popularity scores. This layout allows for easy comparisons between genres and provides a common scale that increases accuracy. The visualization allows users to easily identify the genre with the highest average popularity by sorting the bars based on their popularity values. The highest-rated genre is also highlighted with a different color hue, this is a popout that makes it easier to spot. The visualization utilizes proximity as the genres being compared are situated close to one another. With uniform bin widths and a limited color palette and shape variation of

less than five, the data is easy to perceive from the visualization. Furthermore, due to the utilization of position and color hue, the visualization is entirely visually separable. In this task, we performed discovery and present, exploration and comparison.

Music industry professionals, including music producers, artists, and record label executives, can benefit from this visualization by gaining valuable insights into which genres are currently popular and potentially profitable. This information can inform decisions about marketing, production, and signing new artists.

For duration_s, within 60s, 60s - 120s, and 120s - 180s, which has the highest median danceability score?



This final sketch for visualizing “the highest median danceability score among different song durations (1 minute, 1- 2 minutes and 2 - 3 minutes)” was chosen with the data abstraction task of browsing.

This graph was made by plotting the duration_s (Duration in seconds) on the x-axis and the danceability on the y-axis. The mark used is line and point, with color and position being the channel. The red lines on the “tapes” represents the median and the box represents the interquartile range of the data, and the whiskers extend from the box to

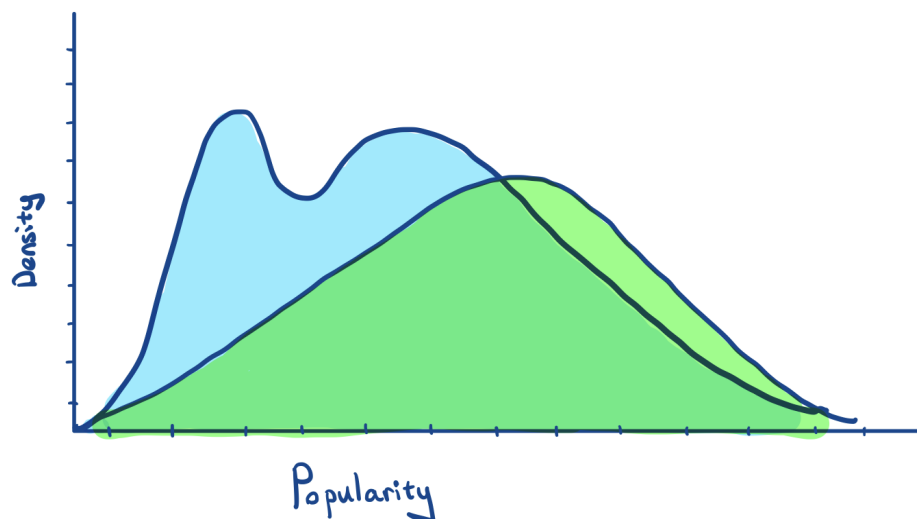
the minimum and maximum values. The use of red color for the median helps the data to “popout”. There’s also “grouping” and “accuracy” in this data that helped make this visualization good.

One key advantage of a box plot is that it displays the median, quartiles, and range of the data, which gives an overview of the distribution. This is particularly useful when comparing multiple groups (classes), as it allows viewers to quickly identify the relative central tendency, spread, and outliers of the danceability distribution.

The Gestalt principle of continuity is used here in the box plot to emphasize the trends and patterns in the data. The median line, in the box, highlights the center of the data, making it easy for the viewer to perceive the overall trend. The principle of closure is also important here with the box and whiskers. This will create a sense of completeness in the plot.

To further improve we can simplify the visualization for higher discriminability.

What is the difference in distribution of popularity between major and minor mode?

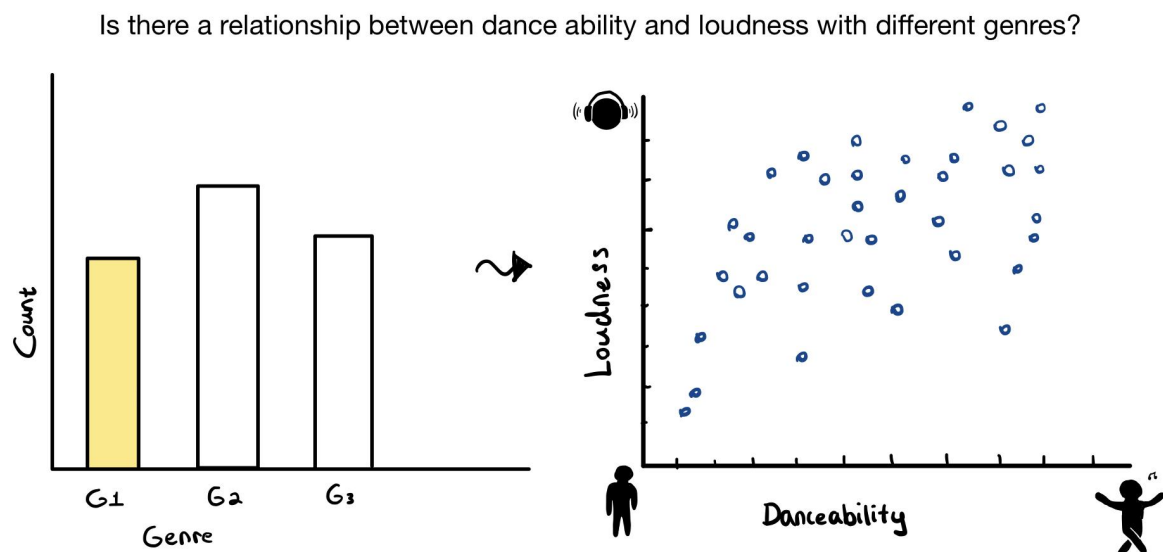


To compare the distribution of popularity between major and minor modes, a density plot would be an appropriate choice. This type of plot displays the distribution of data as a smooth curve, allowing for easy comparison of the shape and spread of the two distributions. By observing the height of the density plot curve, one can determine the

relative frequency of popularity values for each mode, and identify any differences in the density of values between the modes.

The relative frequency of popularity values for each mode can be determined by observing the height of the density plot curve. This can help identify which popularity values are more prevalent in each mode and any differences in the density of values between the modes. In the case of a density plot, the principle of similarity applies because the smooth curves representing the distribution of popularity for major and minor modes are visually similar in shape and form. The plot uses two distinct color hues, one for each mode, which is below the recommended maximum of five, making it easy to interpret without any distortions. The use of distinct color hues for each mode also jump out and allows for quick identification by users. This task can be described as exploration and a comparative analysis of two distributions of a continuous variable (popularity), where the data is grouped by two categories (major and minor mode)

The target audience for this visualization could include movie producers, marketers, and distributors who are interested in understanding which genres tend to be more popular among audiences. It could also be useful for movie critics and enthusiasts who want to explore the popularity trends of different genres over time. Moreover, data analysts and researchers who are studying the movie industry and want to explore the relationship between genre and popularity could benefit from this visualization.



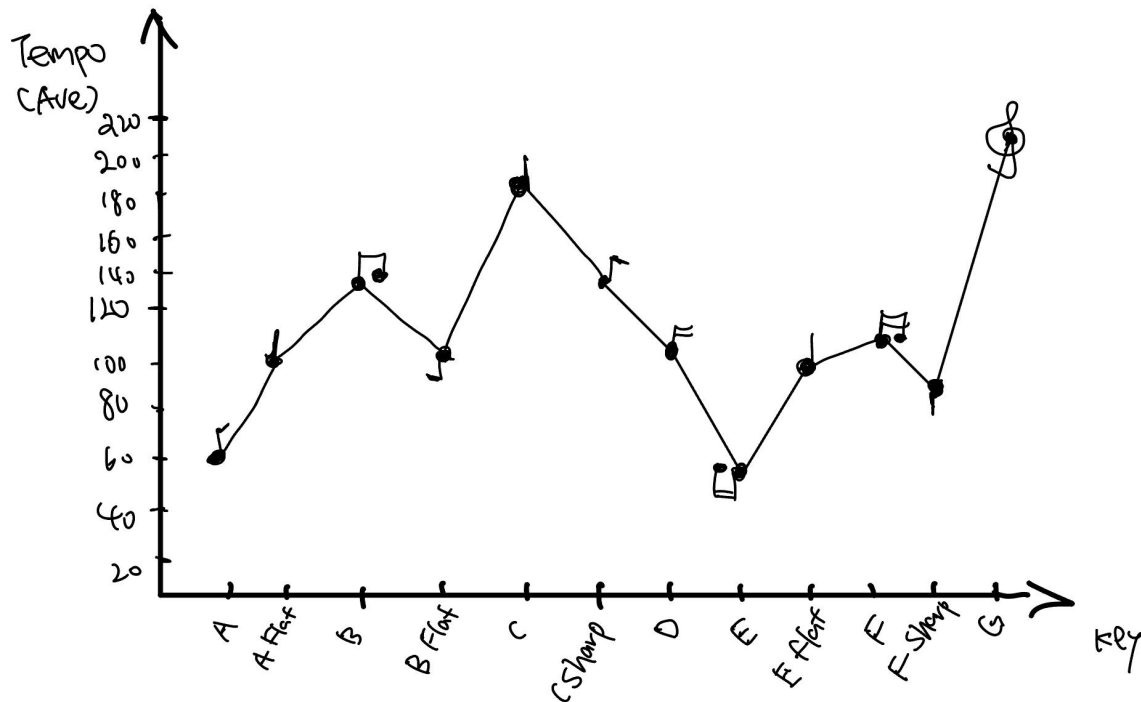
To address the question of whether there is a relationship between danceability and loudness, we have chosen to employ a scatter plot. This type of chart is effective at illustrating the relationship between two continuous variables. Additionally, we have

implemented a unidirectional interaction feature that allows users to view the scatter plot corresponding to a specific genre by clicking on it. This mouse-based interaction is supported by a linking interaction, which filters the data accordingly. We can plot each song's danceability and loudness data points and visually identify any patterns or trends in the data. This makes it easy to see if there is any correlation between the two variables. Moreover, it is very easy to read and interpret, making it a useful tool for communicating the data to a wide range of audiences as it allows people to quickly understand the relationship.

This visualization utilizes a scatter plot with danceability and loudness represented on the x and y axes, respectively. The use of uniform bar widths, limited color hues, and shapes, and a sufficient number of attribute levels ensures ease of perception. The visualization is entirely separable, using position and color hue to differentiate data points. Additionally, selecting a bar highlights it, making it stand out from the rest, further utilizing proximity to enhance perception. This task involves discovery and present, exploration and identification.

The targeted audience may include producers, and label executives. They could benefit from this visualization as it can provide insights into what makes a song popular or successful. Understanding the relationship between danceability and loudness can help in producing and marketing music that appeals to a wider range of audiences. Music lovers may find this visualization interesting as it can help them understand what makes a song danceable and how loudness affects their listening experience.

Is there a difference in the average tempo among each key?



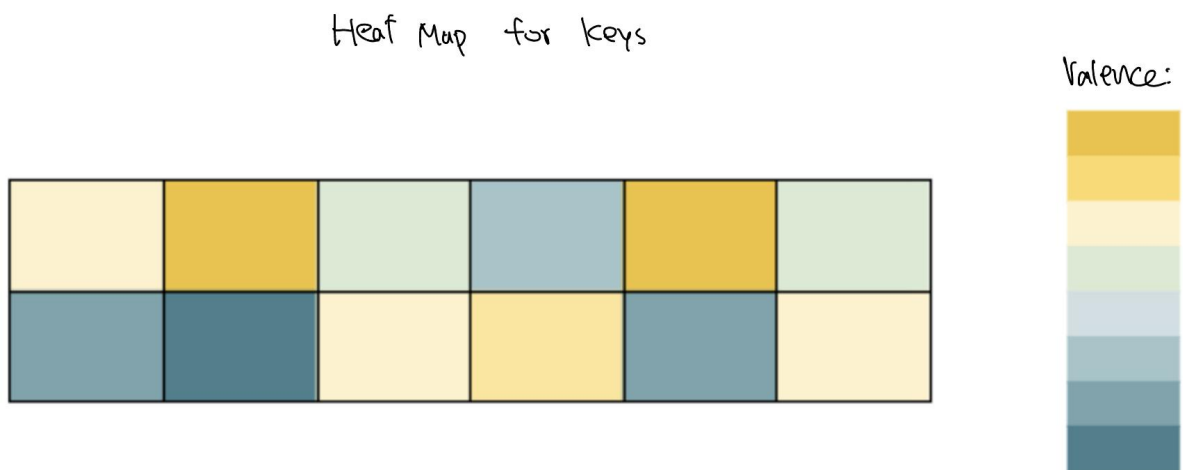
Assuming that 'key' is a nominal attribute, we have chosen this line plot to answer the question: "is there a difference in the average tempo among each key" with the data abstraction task of exploring. A line plot would be a good choice because it emphasizes connectedness and helps us discover trends if there are any existing ones. The visualization also emphasizes "accuracy" with accurate point plotting and "discriminability", with only one step needed to understand data.

We choose to plot the attribute 'key' on the x-axis and the average 'tempo' on the y-axis. The channel used for this plot is position and the mark is points and lines. From this plot, we will be able to identify the exact average tempo # for each key and explore if a change in key (ex. From C to C Sharp) will drastically change the song's danceability.

According to the Gestalt principle of continuity and closure, the lines drawn really create a visual flow for viewers. In addition, the principle of proximity is highly used in this plot, as all the data points are arranged along one line, in close proximity. Both principles used will be making it easier to discover the pattern in data and identify trends.

To make this plot better, we can apply labels and annotations to further help the viewer to identify different data points and gain more information from the visualization. We can also further color the points to increase the “grouping” and “separability” effect for channels.

What's the difference between the average valence # among different keys?



For the last visualization, we used a Heat Map with the data abstraction task of exploring “the difference between the average valence # among different keys”. The channel used is color and the mark is area. We choose to represent each ‘key’ with different boxes, and colors to differentiate different levels of ‘valence’. The color ‘yellow’ represents musical positiveness while ‘blue’ represents musical negativeness. With the Heat Map, we will be able to tell from the color that certain keys contain music that is more positive, negative, or balanced.

In this case, Heat Map would be a good choice because it not only uses the color channel to visualize data but also cognitively creates emotions for viewers with the use of psychological color perception. The visualization also has great channel characteristics of “discriminability”, “grouping” and “popout”, as we are able to identify any extreme numbers, with popout colors and limited steps to understand data.

Moreover, according to the Gestalt Principles of similarity, the visual similarity of colors is used to convey information about danceability. This will help create a grouping effect where areas of similar values are perceived as a group. Furthermore, with proximity, the data points are arranged in a grid, with similar data points gathered together. It will help the viewers to quickly identify if any area has a high concentration of certain groups.

One improvement that could be done to this visualization will be increased continuity. As the human brain perceives objects as continuous, we can use this principle to smooth out the color transition which will help to create a more cohesive visualization representation. Moreover, the accuracy is limited with the color channel. We can solve the problem by adding a label of the average 'valence' number in each box.