Yeehahn Wang-Liu
Ryan Pascual
3/31/2025
Intermediate Data Programming

How and Why Qualities of "Hit Music" Have Changed

Research Questions

1. How have qualities of hit music such as danceability, energy, acousticness etc. changed from the 1920s to the present? How do these changes compare to the changes in music in general?
2. What qualities of music makes it popular?
3. What musical qualities do Grammy nominees have in common? Are some more prevalent than others?
4. Is there a bias for certain artists when selecting a Grammy winner? Are some artists favored? Are some artists prevented from winning?

Motivation

Music influences culture across the world and often reflects aspects of the culture that the music comes from. Through analysis of popular music and the culture that comes with it, people can discover a new understanding of how the general public feels and resonates with over time. For example, what does a decade characterized as having very danceable music say about the culture of the time? While it may be difficult to draw strong conclusions about culture by only looking at hit music, discovery in trends of music can help jumpstart further research in culture.

The Grammy's are a prestigious award ceremony that awards the "best" song with a Grammy award. Winning a Grammy can lead to a "Grammy bounce" where the recognition causes an increase in concert sales, streaming listens, and an overall increase in popularity of the artist. Due to the importance of this award to many artists, many artists have been accused of making their music in such a way it is more likely to win a Grammy but it is unclear what makes a song likely to win a Grammy. There have also been

discussions of the Grammy's being "rigged" or that they only select certain artists and ignore others.

Raw Dataset Summary:

Name	Size	Summary
Spotify_dataset	170653 x 19	Each entry in this dataset is a song and the song's relevant information. The time frame of the songs are from 1921-2020. Each entry contains basic information such as song name, artist, year of release, tempo, etc. However, the most important data for our project will be Spotify's ML generated measurements of certain qualities of songs. Reference *table 1.1* and *table 1.2* to see all relevant columns and descriptions. *(Note: this data set does NOT contain the genre column)*
data_by_genres	2973 x 14	Each entry in this dataset is a genre of music and all of Spotify's ML generated characteristics of that music associated with the genre. Reference *table 1.1* to see ML generated columns.
artists	1162095 x 5	Each entry in this dataset represents an artist on Spotify. It contains general information on the artist and what kind of music the artist creates. See *table 1.3* for more detail.
Grammy_award_data	4810 x 10	This dataset contains all Grammy awards and nominations given from 1966-2025. See *table 1.4* to see all relevant columns.

Processed Dataset Summary:

Name	Size	Summary
artists	6022 x 15	Each entry in this dataset represents a *popular* artist on Spotify. It contains general information on the artist, what genre of music the artist creates, and what the characteristics of the artist's music are. See *table 1.3* and *table 1.1* for more detail.
data_by_genres	2974 x 12	Each entry in this dataset represents a genre of music and all of the ML generated characteristics. Overall, this is the exact same as the raw dataset but two unnecessary columns were removed. Reference *table 1.1* to see ML generated columns.
grammy_song_char	127 x 16	Each entry in this dataset represents a song that earned a grammy and all of the Spotify ML generated characteristics that the song has. See *table 1.4* and *table 1.1* to see all relevant columns. *(Note: This dataset does NOT contain the genres column)*
grammy_award_data	4810 x 5	This dataset contains all Grammy awards and nominations given from 1966-2025. This dataset is a cleaned version (removed unnecessary columns) of the raw one. See *table 1.4* to see all relevant columns.
spotify_dataset	76482 x 14	Each entry in this dataset is a song and the song's relevant information from the time frame of 1921-2020. Each entry contains basic information such as song name, artist, year of release, tempo, etc. However, the most important data for our project will be Spotify's ML generated measurements of certain qualities of songs. All songs in this dataset are loosely considered *popular* by using the popularity column. Reference *table 1.1* and *table 1.2* to see all relevant columns

		and descriptions. *(Note: this data set does NOT contain the genre column)*

Column Name	Data Type	Description
Acousticness	**Float**	A confidence measure from 0.0 to 1.0 of whether the track is acoustic. 1.0 represents high confidence the track is acoustic.
Danceability	**Float**	A measurement of how easy it would be to dance to this song. This measurement is based on factors such as tempo, rhythm stability, beat strength, and overall regularity.
Energy	**Float**	Energy is a measure 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. Perceptual features contributing to this attribute include dynamic range, perceived loudness, timbre, onset rate, and general entropy.
Instrumentalness	**Float**	Predicts whether a track contains no vocals and represents confidence from a scale of 0 to 1. "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. Values above 0.5 are intended to represent instrumental tracks, but confidence is higher as the value approaches 1.0.
Liveness	**Float**	Detects the presence of an audience in the recording from 0 to 1. 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.
Loudness	**Float**	The overall loudness of a track in decibels (dB). Loudness values are averaged across the entire track and are useful for comparing relative loudness of tracks. Loudness is the quality of a sound that is the primary psychological correlate of physical strength (amplitude). Values typically

		range between -60 and 0 db.
Speechiness	**Float**	Speechiness 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.
Tempo	**Float**	The estimated BPM of the song. Most song's typically fall within the range of 60-200 BPM but going outside of this range is not extremely rare.
Valence	**Float**	A measure 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).
Popularity	**Int**	A value between 0 and 100 where 0 is not popular and 100 is very popular. This number is calculated by an algorithm that considers the number of listens and how recent the listens are.
Genre	**String**	A word or phrase that categorizes music based on shared traits or conventions. Spotify's auto-generated and assigned genres are extremely specific and typically do not assign songs to broad genres such as "rock", instead they could assign it to a more specific genre such as "rock brasiliense".

Table 1.1: Data that is collected/generated by Spotify.
**Data for these columns are ML generated. Source of descriptions can be found here.*

Column Name	Data Type	Description
Duration (ms)	**Int**	How long the song is in milliseconds.
Name	**String**	Title of the song.
Artist	**List**	Artist(s) name.
Year	**Int**	Year that the song was released.

Fig 1.2 Table: Relevant columns for spotify_dataset

Column Name	Data Type	Description
Artist	**String**	The artist's name on Spotify.
Popularity	**Int**	A value between 0 and 100 where 0 is not popular and 100 is very popular. This number is calculated by an algorithm that considers the number of listens and how recent the listens are.
Genres	**List**	A word or phrase that categorizes music based on shared traits or conventions. Uses spotify's auto-generated genres and assigns multiple to each artist to account for all music they make.
Followers	**Float**	The number of followers the artist has on Spotify.

Table 1.3: Relevant columns for artists

Column Name	Data Type	Description
Year	**Int**	Year the grammy was given.
Category	**String**	What award was given, generally the most important awards are considered the Album of the Year, Record of the Year, and Song of the Year.
Nominee	**String**	Which musical piece of media (song, performance, album, etc.) was nominated for the award.
Artist	**String**	If applicable, what artist the award went to.
Worker	**String**	If the nomination didn't go to an artist or went to a perofrmance that required a staff (dance performance, movie soundtrack, etc.) includes the

		names of all people who worked on the piece.
Winner	**Boolean**	A true or false statement if the nominee won the award or not.

Table 1.4: All relevant columns for Grammy_award_data

Dataset Challenges

- **General Issues**
  - The “characteristics” of the music are generated by a ML algorithm. Since they are generated by machine learning, there will be deviations from the theoretically most accurate measurement of those characteristics which could cause issues in accurately representing the characteristics of music.
  - All of the data is generated and collected by Spotify (except grammy_award_data) , so measurements in “popularity” only account for popularity on Spotify. This can cause issues with older music as there aren't accurate measurements for how popular an old song was during *its* time. However, since Spotify is the most popular music streaming service the data should hopefully still be generally accurate especially for modern music.
  - The “popularity” category used in multiple datasets is affected by how recent the listens are from when the data was collected. This can cause issues with accurate measurements of popularity of songs that have “faded” away. For example, a song that was extremely popular 5 years ago that is no longer as popular may be ranked lower than a song moderately popular currently even though Spotify does have data of precisely how popular the 5 year old song was 5 years ago.
- **spotify_dataset**
  - There is no genres column in this dataset which will force us to try associating an artist's genres to a song’s genre if we want to discover what genre a song belongs to. Obviously, artists create music for multiple different genres so this may end up in over generalizing an artists style if we take this approach.
- **data_by_genres**
  - The genres are hyper specific and break up general genres into multiple smaller ones. For example, while there is a category called “rock”, there is also “peruvian rock” and “rock brasiliense”

    - The measurements of these songs are taken from averaged and generalized from today. It does not consider how songs were 10, 20, or more years ago.
- **artists**
    - There are hundreds of thousands of artists that have very little to no followers and/or no associated genres. This has significantly bloated the file and we will need to clean up the data and remove all of these artists.
    - Many artists have multiple genres associated with them that are extremely similar. We will need to consolidate these genres into one broader genre so that viewers of our data analysis actually understand what genres we are referring to.
- **grammy_award_data**
    - All of the year recordings are one year behind what they should be.

Challenge Goals

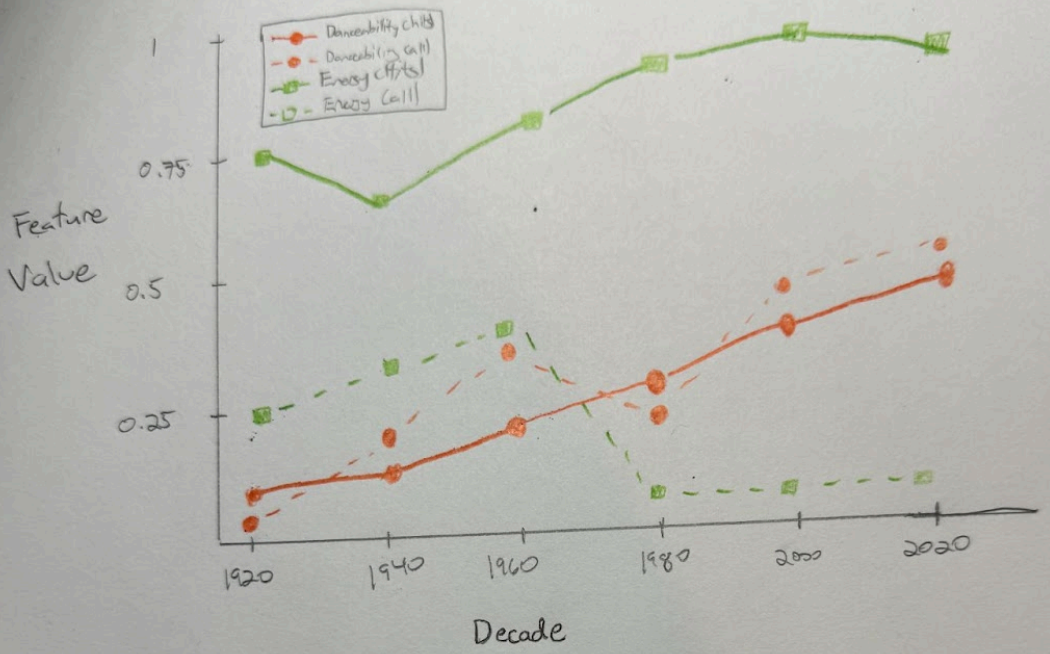
- **Multiple datasets**
    - Most of the data has essentially the same measurements but from different "sources" of music (songs, artists, and genres). We will need to merge and overlap all of these datasets over each other in different ways to collect the full picture. Each of our datasets focus on one specific source of music (songs, artists, and genres) and they are each missing some piece of data that we can fill in by using the other datasets to create a plethora of visualizations.
- **Statistical validation**
    - A chi-square test can be used to help answer research questions 3 and 4 to see if Grammy winners differ from what we expect in a way that is not likely to be chance. However, a challenge in doing statistical validation is determining what we consider to be a "normal" distribution/percentage of grammy wins for an artist. This will help validate our results for questions 3 and 4 in a more concrete way than inspection or plots.

Hit Music Analysis Graph Sketches

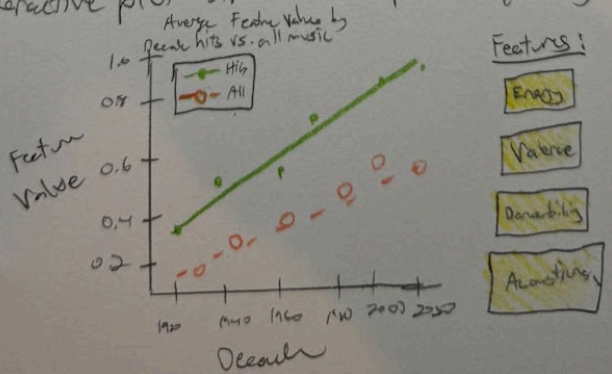
Ryan Pascual
Yeehahn Wong-Liu

RQ #1:

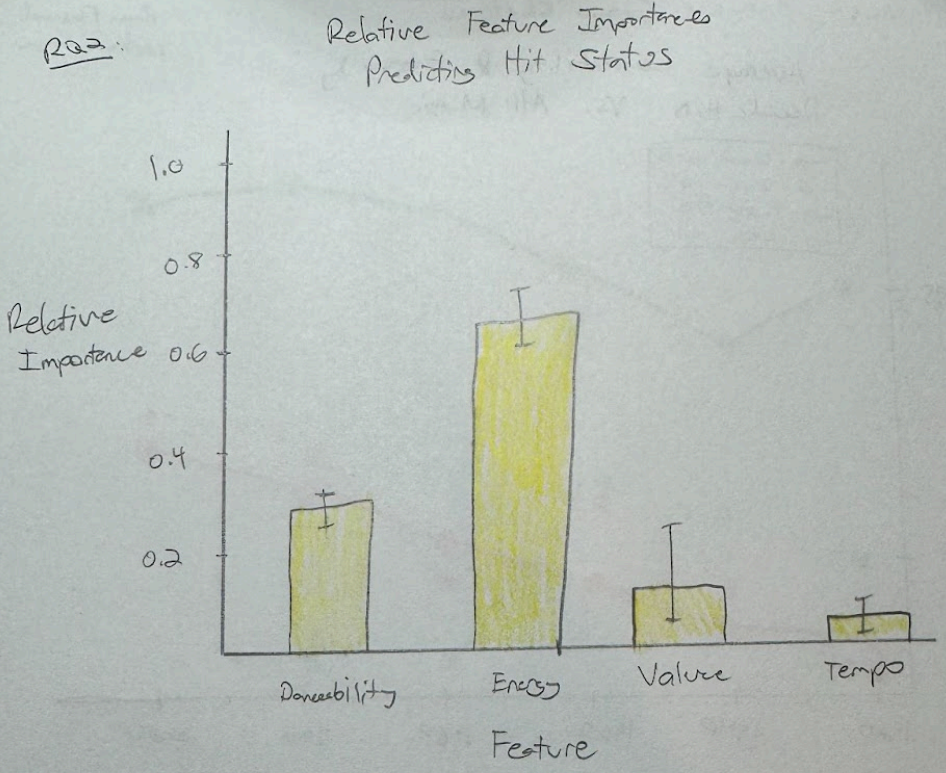
Average Danceability & Energy by Decade Hits vs. All Music



annotations: Can use other musical qualities
Can use more musical qualities
Will pick qualities that paints the most vivid picture
If enough musical qualities are important, can use interactive plot where user picks quality:



RQ2:



Annotations: relative importance is a value from 0.0 to 1.0, obtained from machine learning

depends on how hard it is to obtain, may equally for ML challenge so

Standard error bars represent ±2 SEM

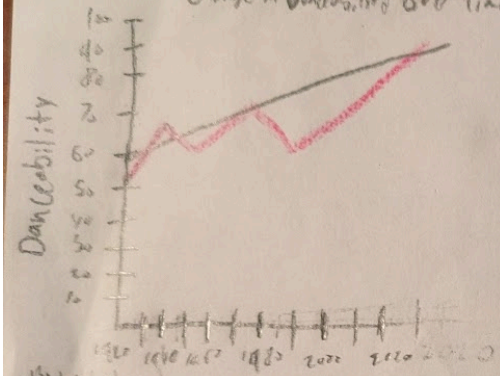
Question 1

- Line Regression Plot
- Looking for strong trends in music

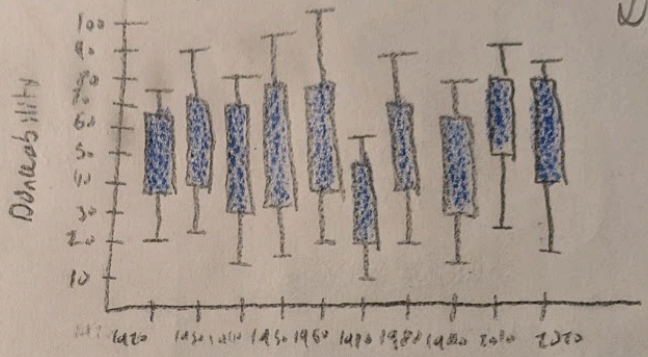
Each graph will be repeated for all applicable ML characteristics. However, for these sketches the graphs will be for danceability, but the idea applies for all.

- Box and Whisker Plot
- Looking for changes but considering extremes as well

Change in Danceability over Time

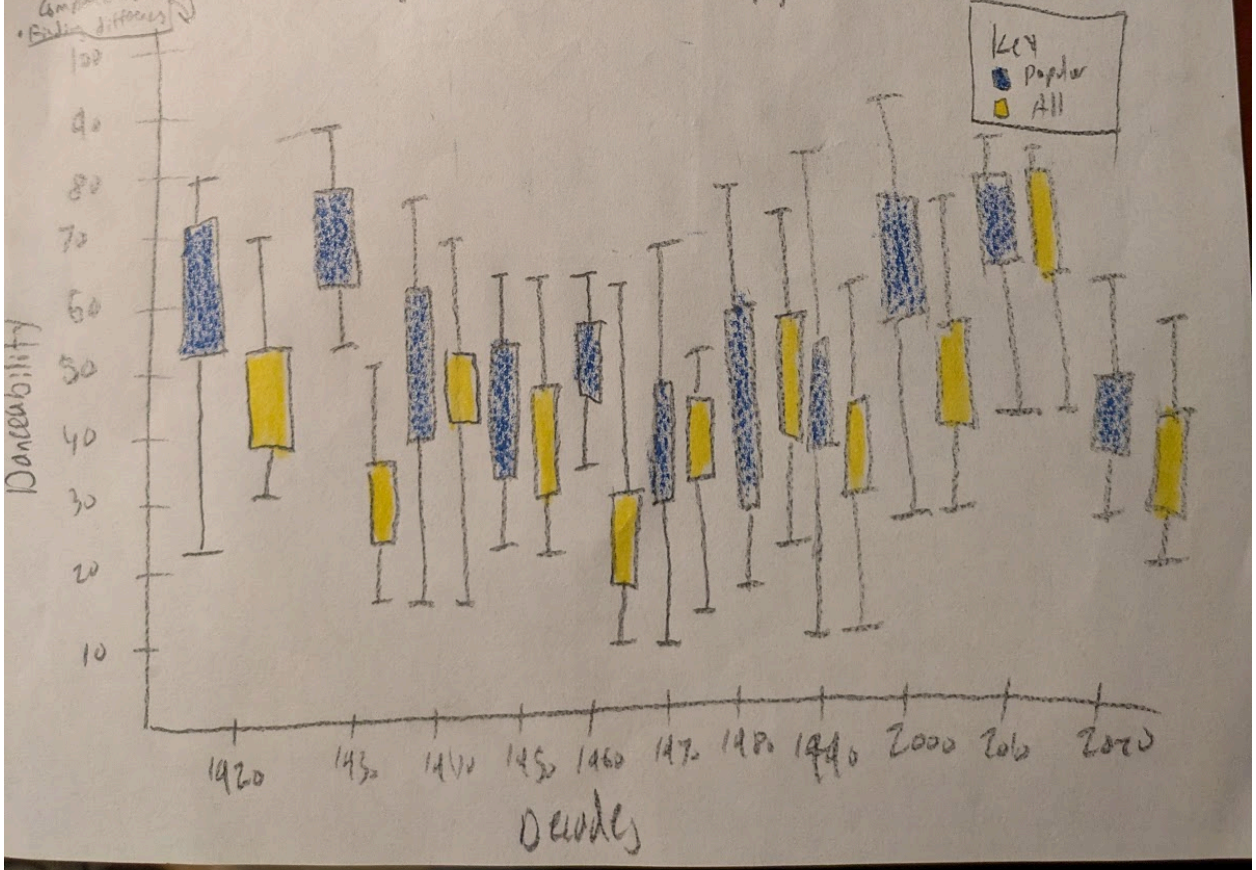


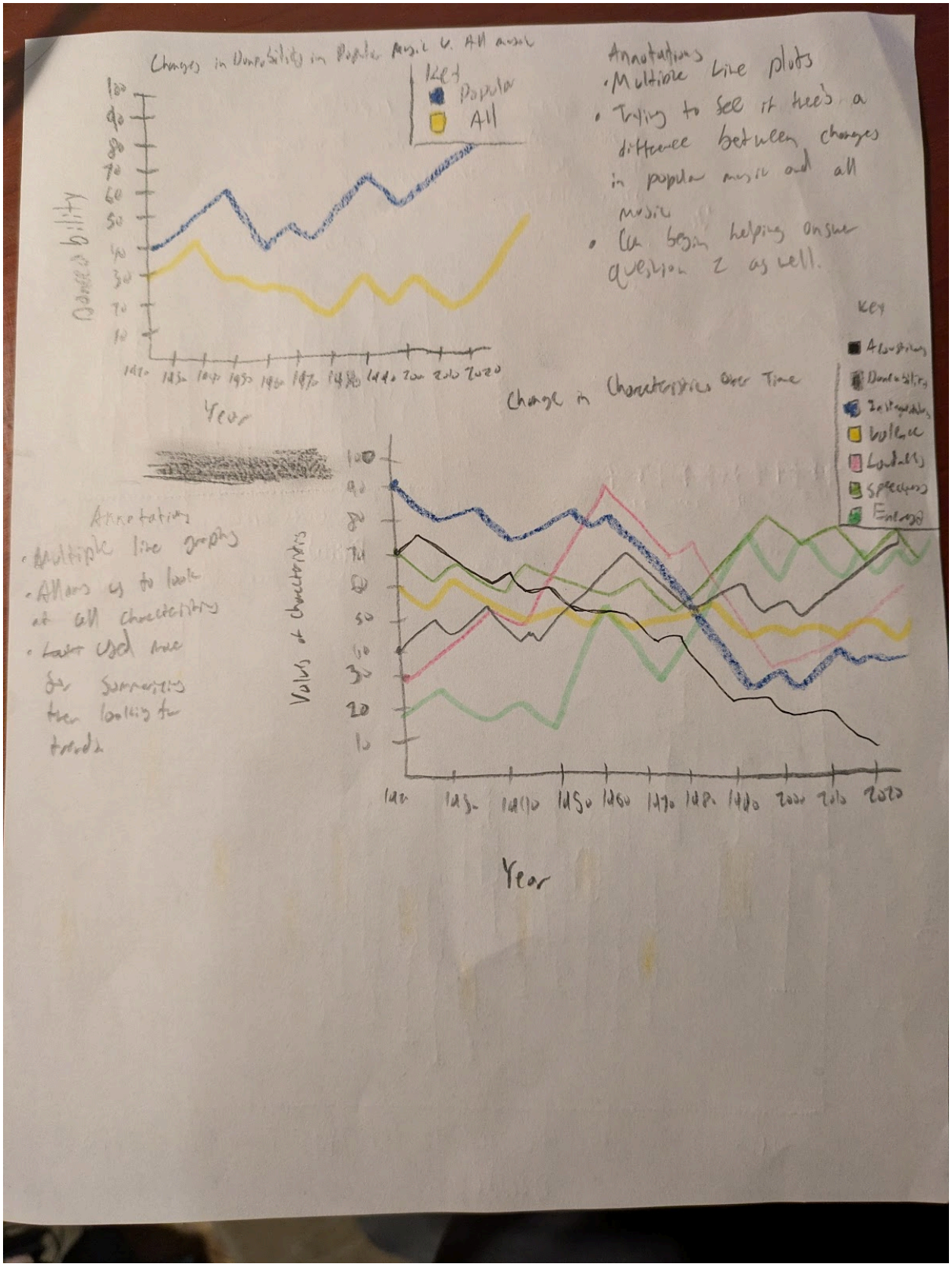
Change in Danceability over Time



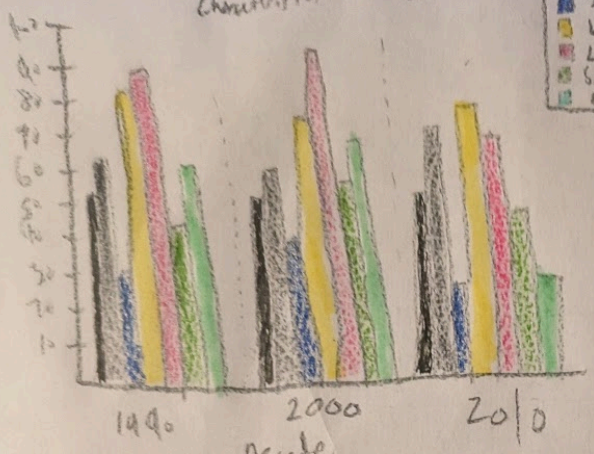
- Box + Whisker Plot
- Trying to compare/contrast
- Bolding differences

Changes in Danceability in Popular music Vs. All music




Changes in Danceability in Popular Music & All music
Key
Popular
All
Danceability
Year
Annotations
• Multiple Line plots
• Trying to see if there's a difference between changes in popular music and all music
• Can begin helping answer question 2 as well.
Change in Characteristics Over Time
Key
Values of Characteristics
Year
Annotations
• Multiple line graphs
• Allows us to look at all characteristics
• Later used more for summarizing than looking for trends

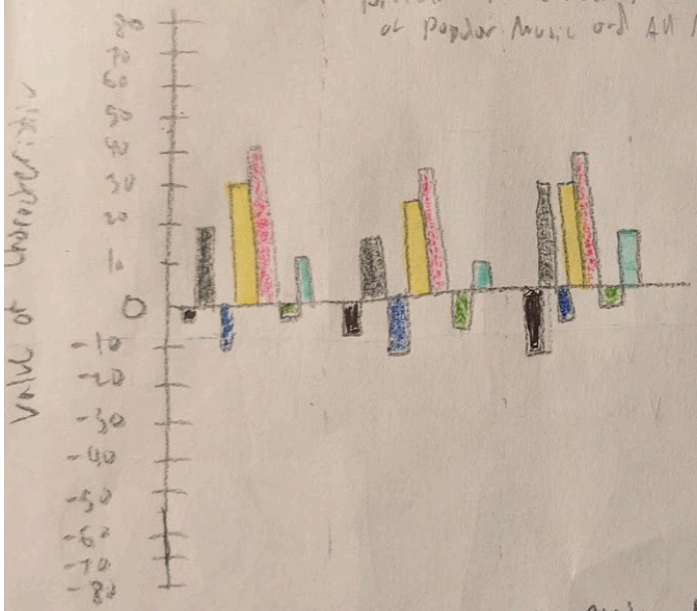
Question 2



Annotations
- We're looking for what characteristics are the strongest
- Multi-Histogram
- Could make multiple plots for more decades but for the sake of demonstration just did 1990 - 2010.

1990 2000 2010

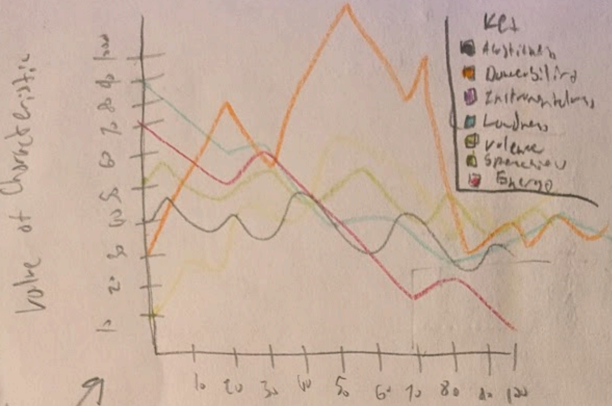
Difference in Characteristics of Popular Music and All Music



Annotations
- Multi-histogram
- We're looking for what characteristic differences are the strongest
- This will help us find what makes music popular what the difference is

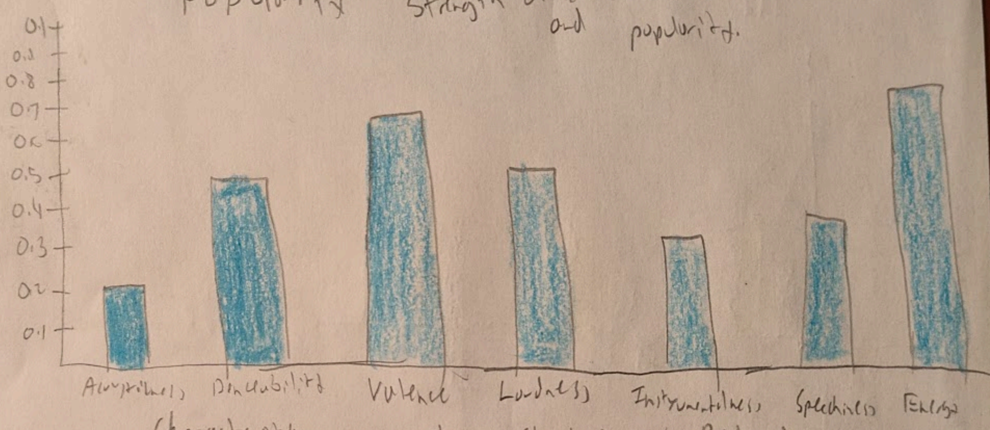
Characteristics of Popular Music - All Music

Traits vs. Popularity



- Multiple-Line Charts
- The data displayed in this graph can be used to find correlation between characteristics and popularity
- While there is the option of splitting this into multiple graphs the main purpose of this graph is to further help display

Strength of Correlation of each characteristic and popularity.

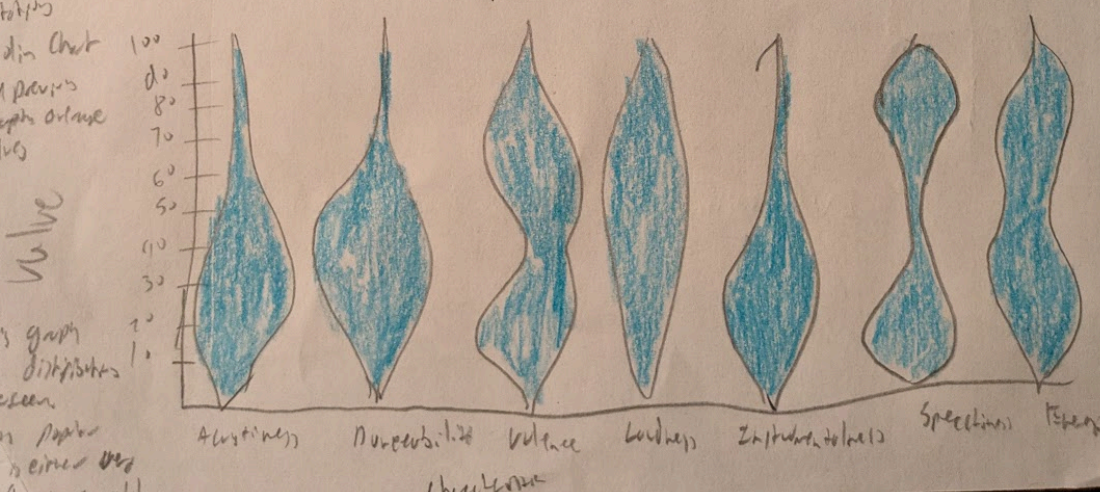


- Bar Chart
- R^2 value will be found from
- The higher absolute value the stronger the correlation
- Help discover what creates the most popular music

Anotatyps

- Violin Chart
- All previous graphs' average values

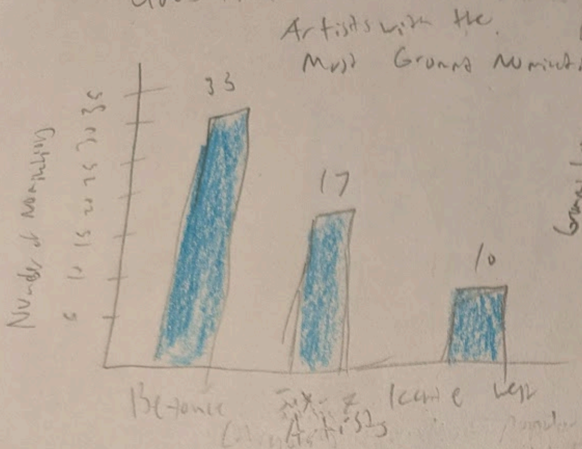
Violin Plot of Characteristics of Popular Music

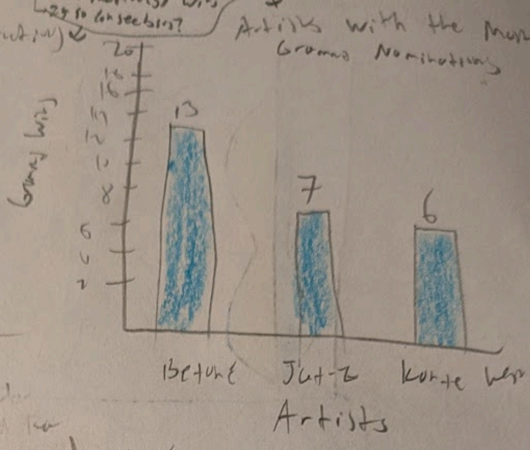


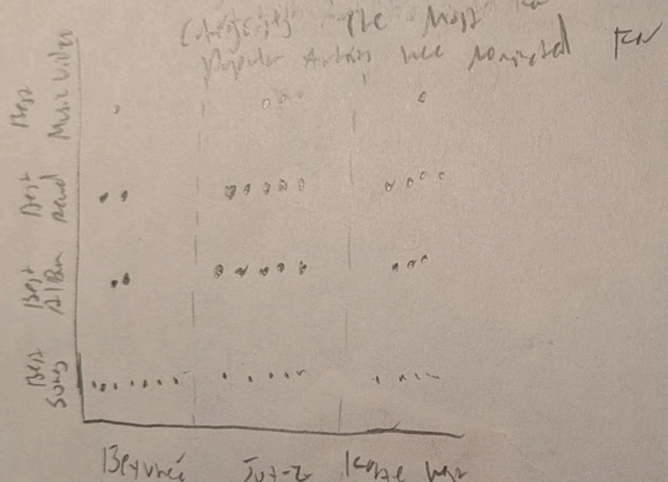
- This graph allows distributions to be seen.
- Perhaps popular music is either very little ... or alt.

Question 4

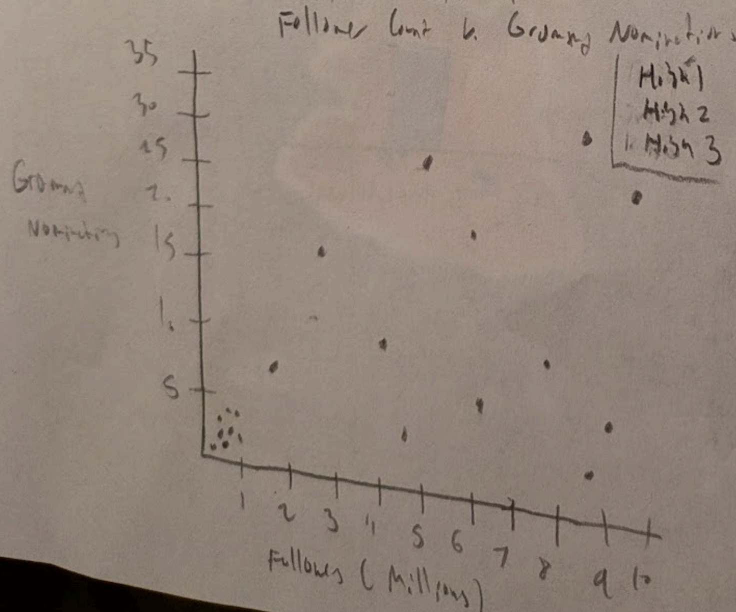
• Bar Charts
• Looking for artists with the most nominations/ wins







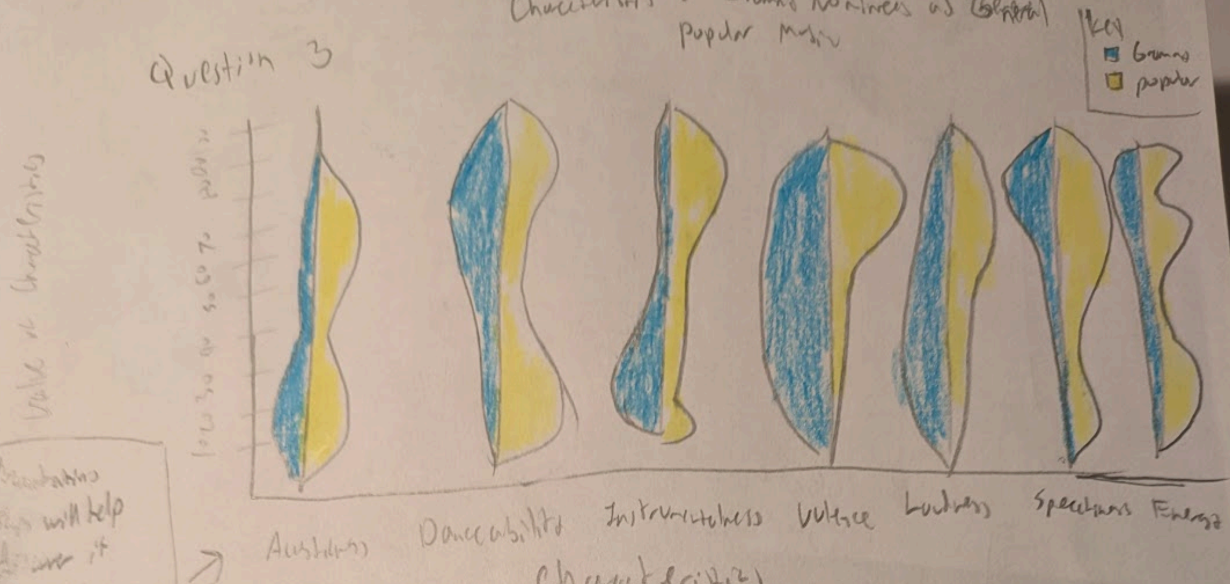
• 2D dot plot or swarm plot
• Trying to see if there are any categories with favorites



• Looking to see if there are artists who have many nominations even though popularity may not reflect that.

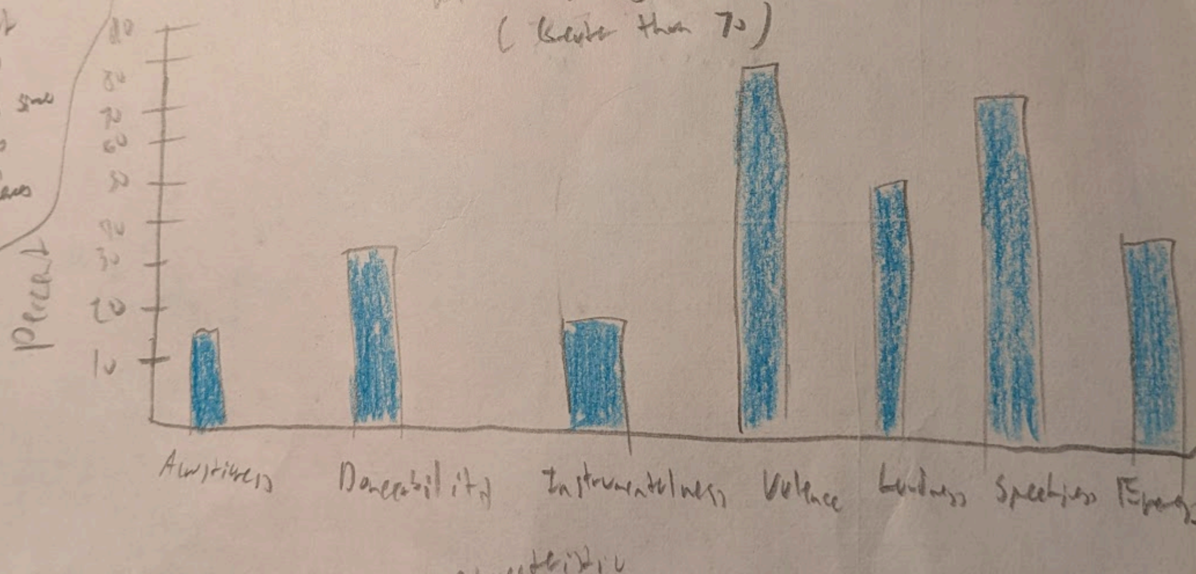
Question 3

Characteristics of Grammy Nominees and (General) popular music



Annotations will help discover if there is a difference in Grammy songs and popular songs.
- Violin Chart
- Ensures everything shows distributions are shown and differences clear

Percent of Grammy Nominees with a strong characteristic (Greater than 70)



Annotations
- Will help show if Grammy songs have a tendency to be high in a characteristic
- Bar Chart