**Analysis Summary:**

**Subject of the study**

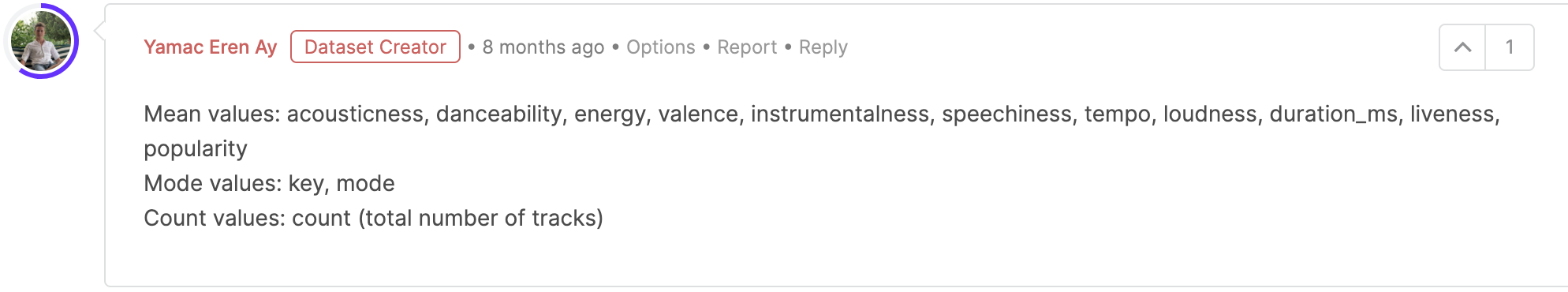
We analyzed Spotify data spanning the years 1920 to 2021. The data sets included facts and song characteristics, such as artist, music genre, song popularity and music characteristics. The goal of our project is to look at the dataset from a high level and uncover music trends. We are interested in how artists, genres and popularity of songs developed through time.

**Major Questions/Goals of the Study**

1. Conduct a high-level descriptive analysis of the entire dataset
2. Determine the most timeless artists
3. Determine if there are significant trends in song or music attributes over the years
4. Determine if any music attributes are less or more significant today vs the past
5. Understand the number of songs uploaded to Spotify each year
6. Calculate correlations amongst the song characteristics as well as popularity vs the song characteristics
7. Determine if one or two audio variables can predict a song’s popularity
8. Determine if genres have a defining set of music characteristics and if commonalities between genres exist

**What methods were used to answer the major questions?**

1. Data Collection: Four csv files were collected from Kaggle ([link here](https://www.kaggle.com/yamaerenay/spotify-dataset-19212020-160k-tracks?select=data_by_artist.csv)) and three were utilized for the study
2. The datasets were inspected and cleaned prior to analysis. There were no missing values, the data types are appropriate (number fields are integers or float, text fields are objects. Three of the datasets were data that are “groupby’ed” and summarized with mean values so very little cleaning needs to be done for our analysis. The creator of the datasets have done the leg work for us.
   1. The exact way of how the data is collected is not known, but below are some additional evidence to help guide us:



1. Analyses included:
   1. YoY Analysis - time series (1920 - 2021)
   2. Correlations and average values for:
      1. Acousticness
      2. Danceability
      3. Energy
      4. Duration\_ms
      5. Instrumentalness
      6. Valence
      7. Popularity
      8. Tempo
      9. Speechiness
      10. Liveness
      11. Loudness
2. Data was analyzed via dataframes and groupby’ed where needed
3. Line plots and histograms were utilized to show trends in music characteristics
4. Radar plots were chosen to visualize genre-specific music characteristics
5. Relationships between popularity and music variables were shown using scatterplots
6. A half-matrix heat map was used to visualize correlations amongst music characteristics and popularity

**What are the major conclusions from the study?**

1. Music characteristics appear to have evolved over time and began stabilizing around 1960
2. Major genres have similar music characteristics like tempo, but some divergence noted for explicit and accousticness variables
3. No single music characteristic can predict a song’s popularity; some covariance found amongst the characteristics
4. Multiple linear or nonlinear regression would be an interesting next step towards predicting music popularity but not explored.
5. Songs from the late 1990s to the early 2000s are the hottest on Spotify.
6. Music’s legends continue to be popular today

**What were some of the limitations of the data set(s)?**

1. Lists of collaborators such as (['JAY-Z', 'The Notorious B.I.G.']) makes analysis relevant to individual artists tricky
2. Genres were loaded as lists of values and are sorted alphabetically, making it hard to determine popular genre trends
3. Popularity is a subjective measure and a variable. It’s based on episodes of listening by contemporary listeners at the time of data collection, and it will vary over time based on Spotify users’ listening patterns.
4. Data collection methodology not exactly clear, especially when accounting for noisy data in early 20th century

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

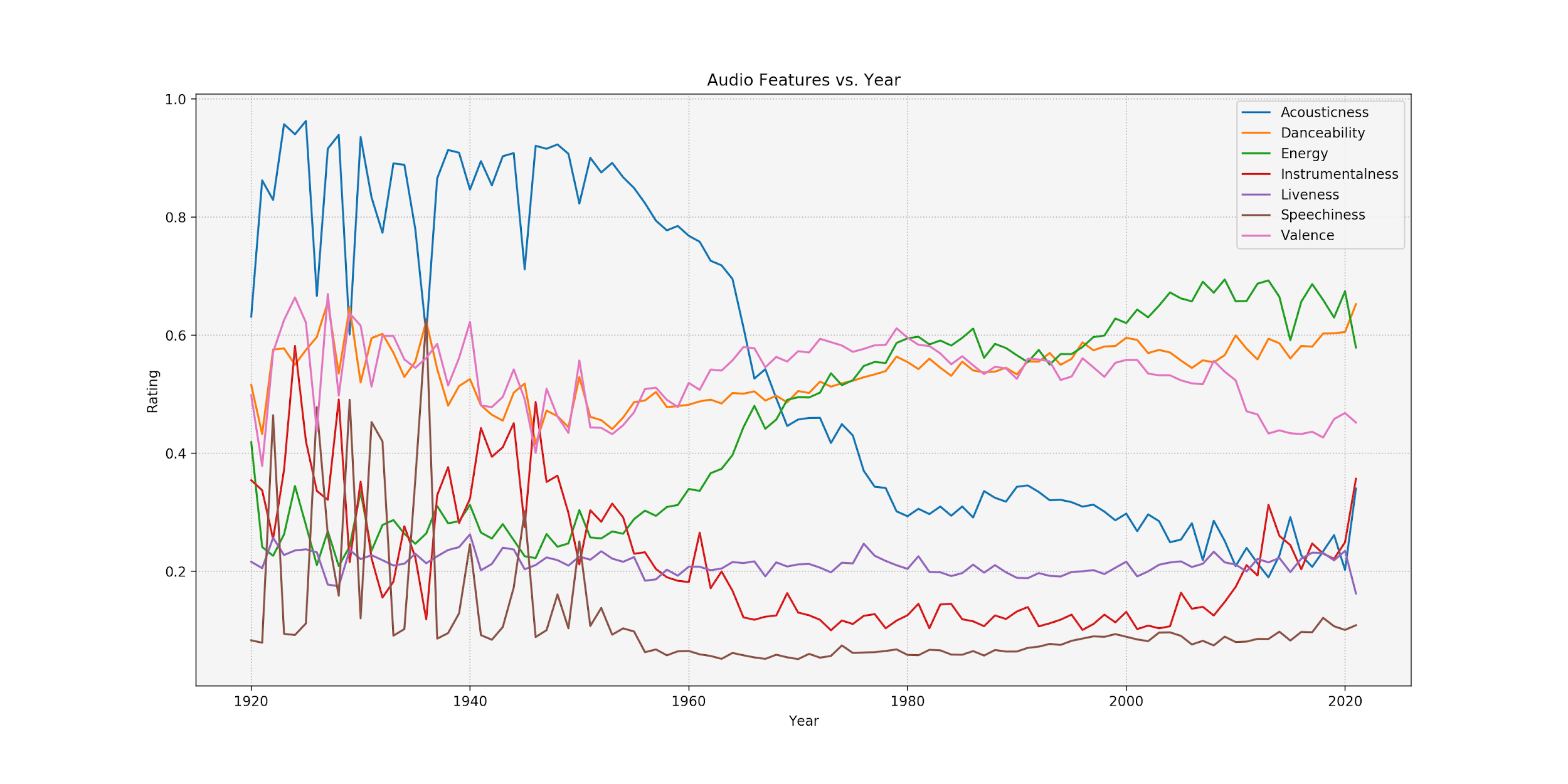
**Summary discussion from each team member’s contribution on analytical questions**

**Jeff: How has Spotify data developed over the years?**

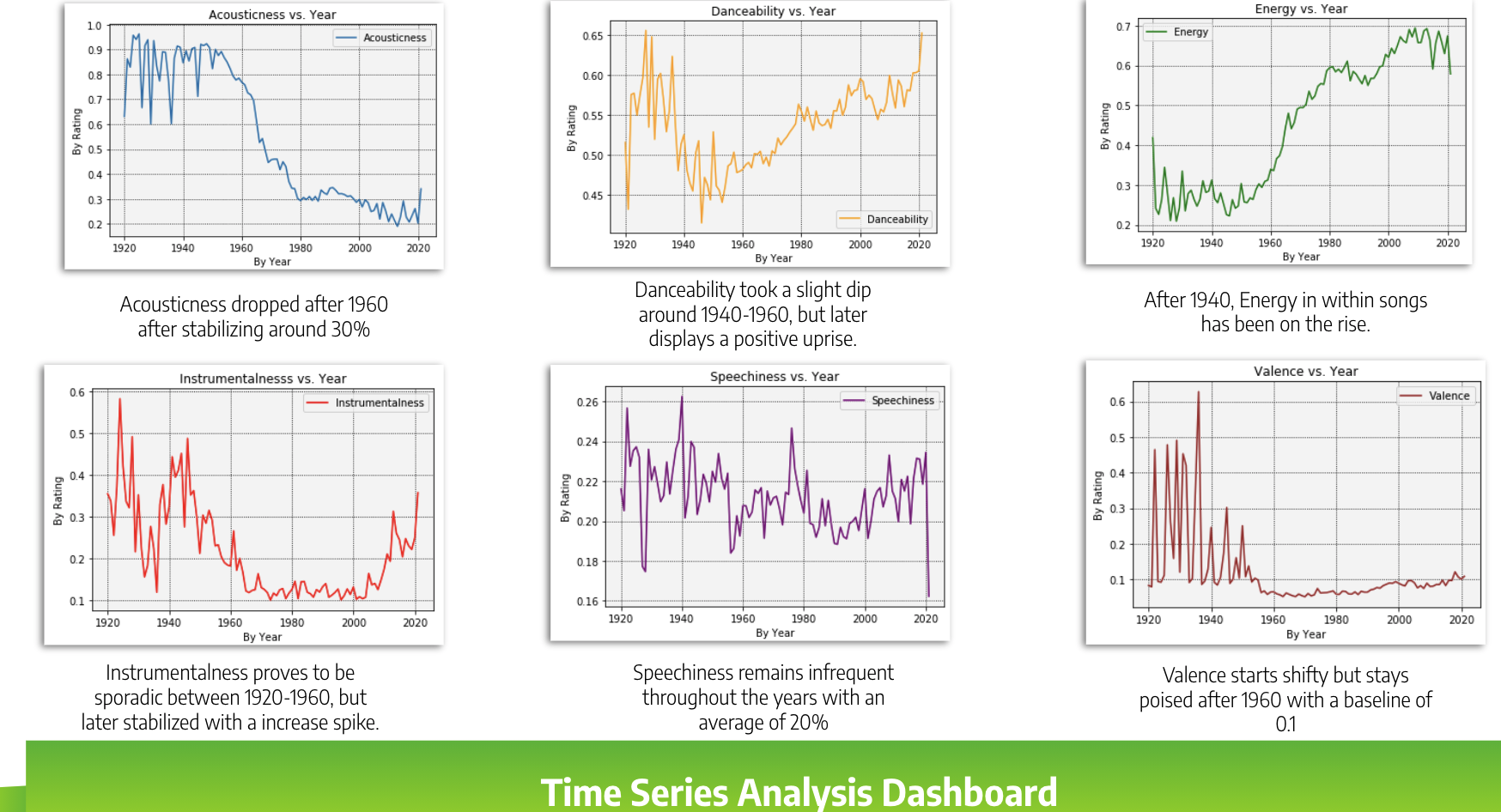
Analysis file: time\_series\_analysis.ipynb

Approach: Spotify trends over the years is one of our developing questions. We decided to create a multi-line graph to capture the relationship between the dependent variables. The x-axis highlights the timeframe between 1920 through 2021 whereas the y-axis is the different characteristics by ratings.

Findings:



Dashboard:



Summary:

A final analysis of the below charts, the data is very sporadic from 1920 - 1980, and then later stabilized. Which led our group to question the shift of frequency around 1980. Therefore, we came up with a few interesting factors. But to highlight on the top two factors which is the period behind the digital age and the radio station outreach. We believe the digital era and radio station influence has enhanced the music industry to cater to a much larger audience when it comes to genres and taste . Versus before 1970 the sample size was much smaller when it comes to data collection and targeted audience.

**Jeremy: Song Characteristic Trends Amongst Popular Genres**

Analysis file: artist\_genre\_radar.ipynb

Approach

* Genres were chosen by picking artists that best represented each genre from the Billboard 200.
* The cleaned\_data.csv data was imported as a dataframe and then inspected to ensure the artists were both contained in the data set and had at least twenty songs for the analysis (using ‘count’ data column).
* The dataframe was then reduced to contain only the artists and columns of interest:
* Artists: Metallica, Avicii, Eminem, Taylor Swift and Blake Shelton
* Relevant data columns: Acousticness, Danceability, Energy, Duration\_ms, Instrumentalness, Valence, Popularity, Tempo, Speechiness, Liveness, Loudness and popularity
* The data set was sorted by artist and song popularity and then the dataframe was modified to only the top 20,most popular songs for each artist
* Most song characteristic data had values between 0 and 1; however, tempo, loudness and duration had data ranges that were higher or lower than 1. We employed normalization on these three data columns to better visualize all of the song characteristics on one graph.
* A simplified data frame was then generated via groupby artist and mean() method.
* Histograms for each artist and each characteristic were inspected to ensure the average of a song characteristic was a suitable representation for the artist
* Lastly, the average song characteristic for each artist was plotted on individual radar charts and then overlaid on one radar chart for comparison.

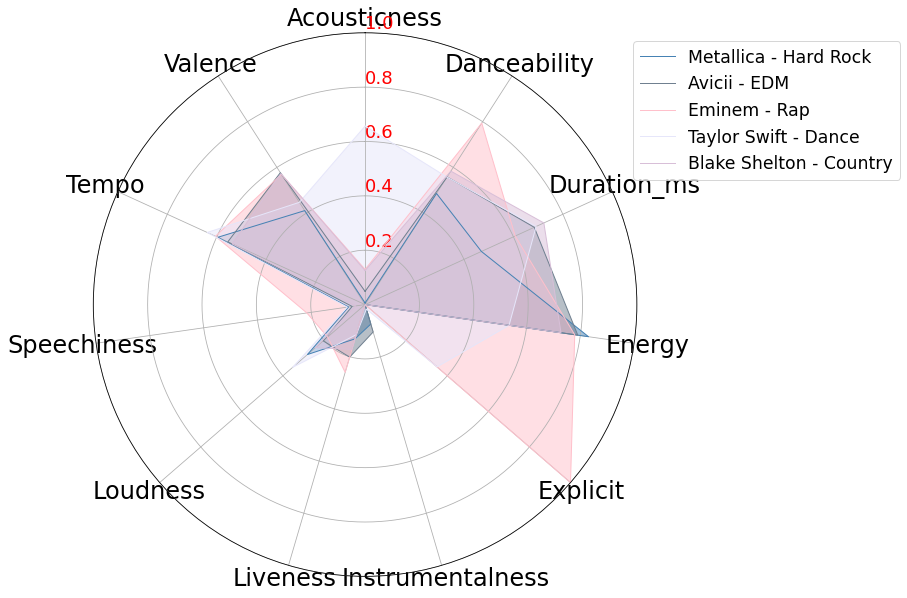
Major obstacles

As mentioned previously, Spotify’s data files and API associates each artist with several major genres and sub-genres. The genres are also listed alphabetically vs most relevant or most listened to or most requested, which further complicates any analyses to determine music trends relating to artists. The group initially wanted to concentrate on the most popular genres of today, which was decidedly difficult with the chosen data set’ structure.

Data Gathered

All plots and data sets are shown in the above mentioned jupyter notebook; only the overlapped radar chart is shown here for brevity and because this chart represents the sum total of the analysis into music genres.

Music characteristics are shown along each angle and values for each characteristic emanate from the center out to their individual, averaged values. Conclusions from the study are summarized below.



Major conclusions - Genres and Music Characteristics

* Major differentiation amongst explicit and accousticness categories for at least two of the artists
* Some delineation between valence and energy amongst the group, possibly an indication of a genre’s fingerprint
* Alignment or agreement amongst the genres for tempo, loudness, liveness & danceability
* Live recorded music doesn’t appear to be popular option for these genres
* All genres chosen have low speechiness and low instrumentaliness
* One could almost argue the almost wide-spread agreement or overlap of characteristics amongst the genres could indicate society is influencing this overlap or the music industry is relying on specific formulas per se and manufacturing today’s music to ensure popularity

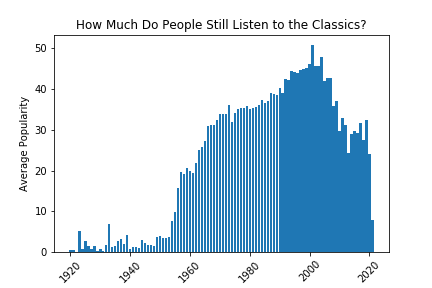
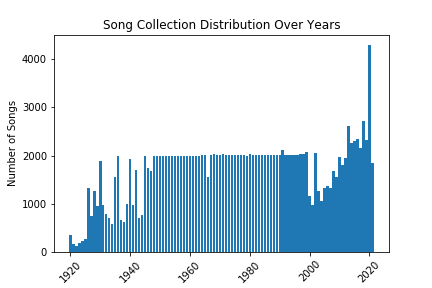
**Tianchi: (1) Spotify User Music Taste and (2) Music Goat Discussion**

Analysis file: exploration\_descriptive\_analysis.ipynb

Complimentary file: data\_gen.py; animate\_time\_series.py

Approach (1) - Spotify User Music Taste:

We took the number of songs released each year to further understand our dataset. It was also necessary to have that knowledge first before we explore which era are the songs most popular in contemporary standards because data collection methods could skew the variables in unnecessary ways. We then took the average popularity score for songs from each year to analyze potential trends.



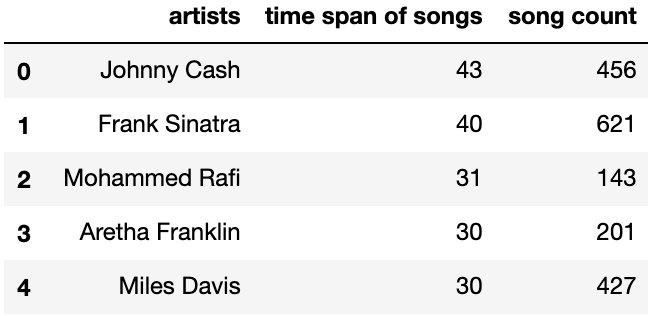
Findings (1):

Looking at the grah on the left (Song Collection Distribution Over Years), the author addressed in the Q&A on Kaggle that he capped the maximum amount of songs collected (from Spotify API) at 2000, yet some of the years have way fewer than 2000 songs and some have more. It’s understandable that some years have fewer than 2000 because there could be only so many songs from a particular year. However, it raises doubts that some of the more recent years (early 2000s) have fewer than 2000 songs. Due to that, we concluded that the dataset creator could have made a mistake. However, since this does not affect our analysis in any catastrophic way, we decided to keep all the data intact and not remove any rows. Over on the right, there is a clear trend that the average popularity increases as the years approach current time. The avg. popularity peeked at around 2000 and started to decline. It is interesting to visualize the trend and we came up with two interpretations/theories:

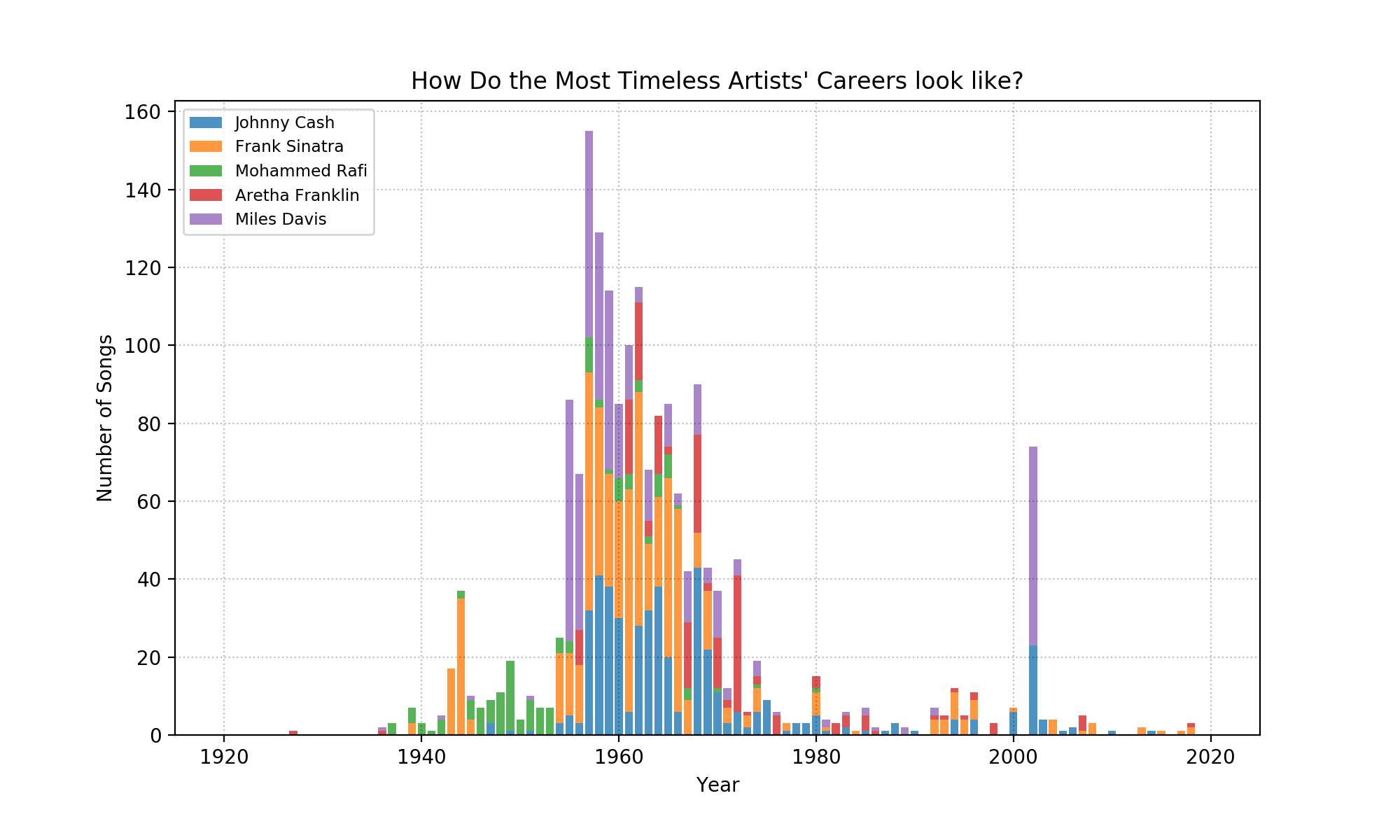
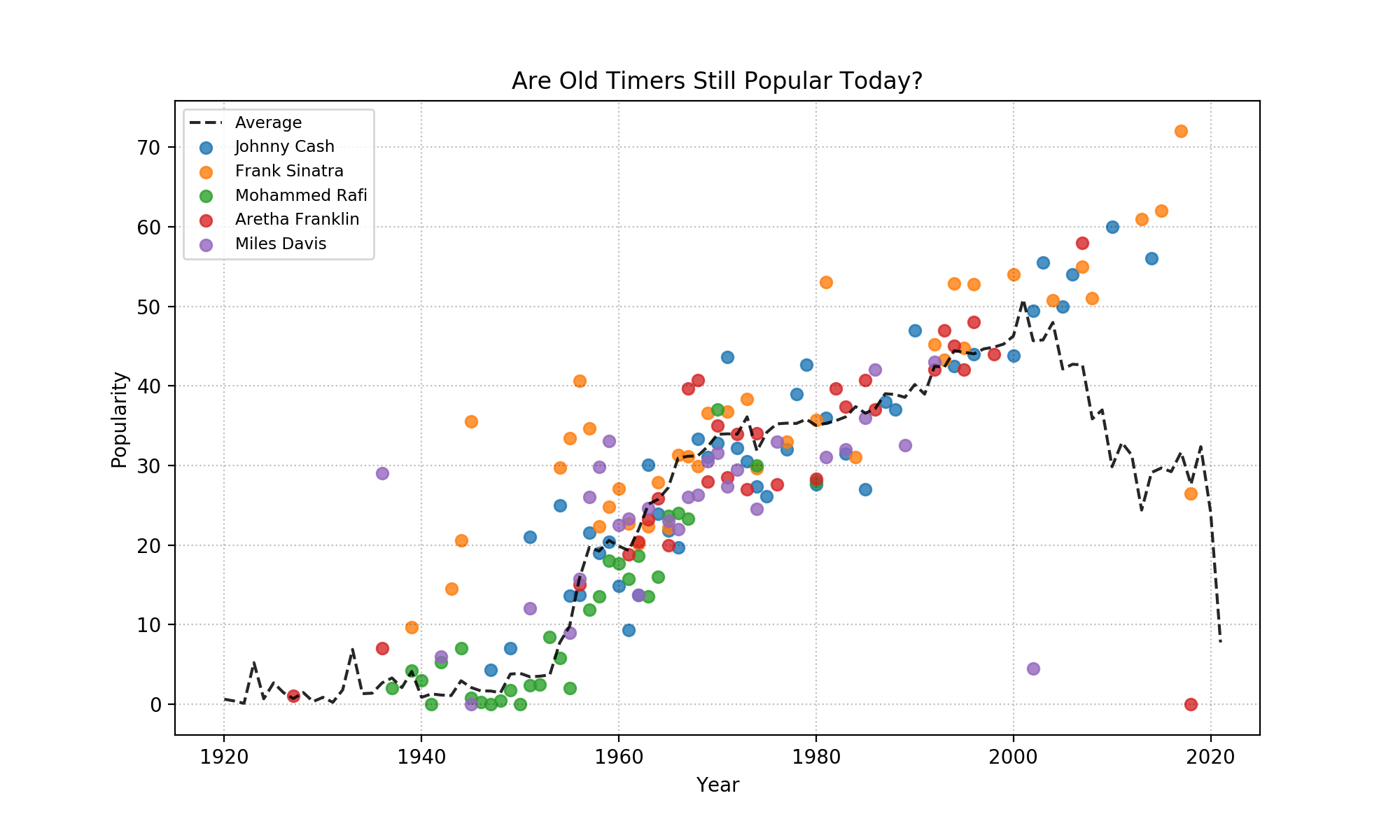
1. Spotify’s user base has people from different generations. Older generations could potentially listen to relatively older songs more frequently and vice versa, young generations listen to relatively more contemporary songs frequently. In fact, “As of March 2018, Spotify's user base was dominated by Millennials, with 29 percent of its users aged 25 to 34 and 26 percent aged between 18 and 24 years old.” ([Statista](https://www.statista.com/statistics/475821/spotify-users-age-usa/#:~:text=As%20of%20March%202018%2C%20Spotify's,18%20and%2024%20years%20old.)) As a result, it’s no surprise that as the years increase, popularity increases since the Millennials are probably listening to more contemporary songs. As to why the avg. popularity are fairly high and well above overall average from the 1960s to the 1990s, our explanation is that people are nostalgic. Which leads to the next theory:
2. As to why there’s a downward trend after 2000: Songs are just like arts, they take time to be appreciated and recognized. That could explain why some of the older songs (from 1960 to 1970 for example) are still fairly popular because people like listening to older songs. So if we could somehow fast forward to 2060, we could possibly (only possibly) see the popularity from 2000 to 2020 start to pick up, and some of the older songs could start to drop as the trend indicated.

Approach (2) - Who are the most timeless artists?:

We wanted to understand how artists’ careers flourish and decline, and we wanted to debate who the most “timeless” artist is. To answer this question, we first had to define what “timeless” means to us, so we decided to look at the number of years each artist’s songs were published across the entire dataset. This sort of defines timeless as longevity. In addition, we also looked at the total number of songs each artist has (at least in this dataset). We believe the song count is also an important attribute and should aid us to make our decision. As a result, we summarized the data and ranked them in descending order (see below). Johnny Cash and Frank Sinatra are the only two that have songs published through at least 40 years and also have more than 400 songs. We will be focusing on these two.



Findings (2):



Looking at the charts above, we plotted the average of popularity and number of songs for each artist by year separately. If we compare the overall popularity trends of these 5 artists, they closely follow the overall average of the entire dataset (black dotted line). Arguments could be made that they are driving the overall average because they each has a large number of songs. However, there are plenty of instances where their popularities are above the overall average. For example, Frank Sinatra’s popularity over the 40s to 60s (which also were the era when he produced most of his songs by looking at the second bar chart) are well above average compared with peers at the same timeframe. Also, even after Frank and Johnny died (in 1998 and 2003 respectively), their songs after 2000 (likely republished or songs collected into new playlists by people online) have popularity well above average as well.

Conclusion:

Time is the greatest testament of someone’s greatness. Even though Johnny and Frank are two close comparison, given he has about 200 more songs and also higher/more popularity distribution from our scatter plot, we conclude that Frank Sinatra is the most timeless artist.

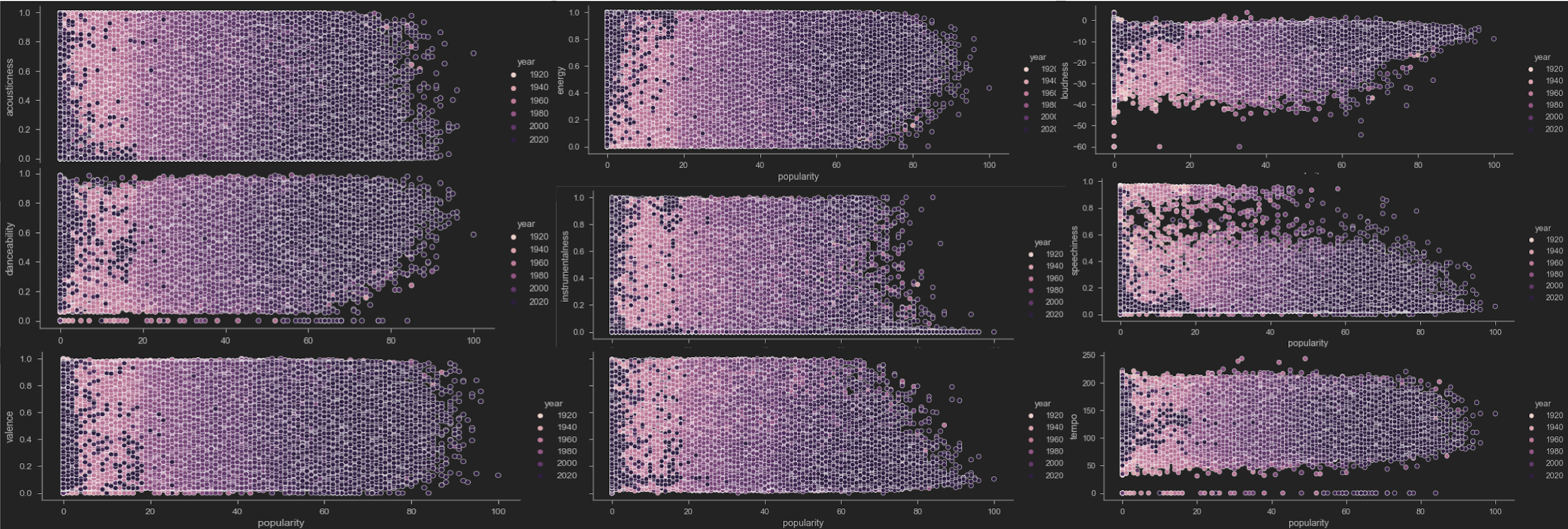
**Aaron: Are there correlations between audio features and Does one or two audio features predict a song’s popularity effectively?**

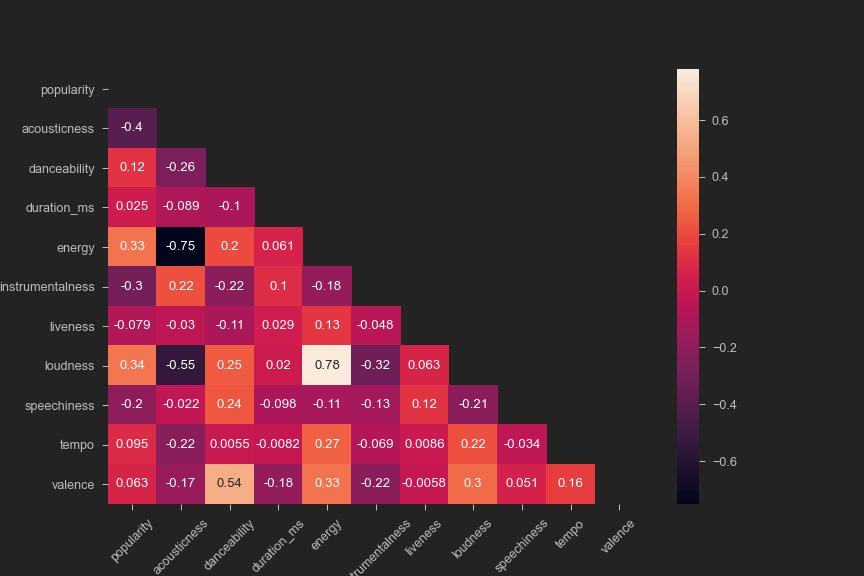
Analysis file: correlations\_analysis.ipynb; number-of-songs-collected-per-year.ipynb

Approach - Pearson Correlation Value Heatmap:

It was really crucial to look at all of the correlations between all the variables with one another, but most importantly, we want to see if there is any correlation between any of the variables and popularity. We created a scatter plot to visualize the data, but since there are thousands of entries, we also created a Pearson Correlation heatmap in order to be able to see the relationships between each variable to one another.

Findings:





Looking at the scatter plot, they are color-coded based on the release year of the song. The darker the color, the later the song is. We do not find any correlation between the year of the song and any of the variables we looked into, since all the plots are scattered evenly by color.

Based on the Pearson Correlation heatmap, the darker the color is, the stronger the positive correlation between the two variables, and the lighter the color is, the stronger the negative correlation between the two variables, whereas the r-value of 0 is represented by the color red, which means that there is no correlation between the two variables. Looking at the relationship between popularity and all of the variables, we found there is a low negative correlation between acousticness and popularity with an R-value of -0.4, and also a low positive correlation between loudness and popularity and energy and popularity with both R-value being 0.33.

When comparing the rest of the variables against each other, We found out that there is a negative correlation between acousticness and energy, and, acousticness and loudness, and also we see a positive correlation between loudness and energy, and danceability and valance.

Conclusion:

Through a correlation of each audio feature against popularity, we found out that there is no single variable that affects the popularity of a song in a major way. Since the correlations between the variables are relatively weak, we believe that a multiple regression wouldn’t really show a very effective way to predict popularity given if we want to consider all of these audio features all together, but it requires further analysis to determine the effectiveness of a model like that. We decided to prevent us from going down the rabbit hole before we find ourselves limited by our skills and knowledge at this moment, so we will conclude.