Analysis of Spotify Songs Features and the K-pop Genre

Authors:

Laphon Premcharoen and Phoenix Yi

Background and Motivation:

Spotify is a popular music streaming application which offers wide ranges of genres with over 60 million songs on the platform. Spotify is available on most digital devices, including personal computers, smartphones and tablets with 299 million monthly active users in July 2020. With Spotify, users can browse songs by artists, albums, playlists and genres, and can also customize their own playlists. Aside from the graphic interfaces, Spotify offers a web API where developers can access its database. Through this web API, developers can obtain audio features of a song including keys, tempo, loudness and duration etc. With its big accessible database with diverse data, Spotify is a suitable platform for song analysis. In this project, we have decided to perform analysis on audio features of Spotify's songs to see the trend of song consumption over time. We are also going to analyze the K-pop genre specifically as it is currently one of the most growing genres as a non-English genre. We are interested to see if this genre has unique features that set it apart from others

Research Questions and Results:

- **Is there any commonality among popular Spotify songs in the year they were released? and has the trend changed over past years?**
    - The mainstream music taste is changing over time. By visualizing the correlation between audio features of the top songs and the years they are ranked, we may see the trend of the mainstream music taste and be able to predict the music trend in the future.
- **How have typical features of songs changed over time?**
    - Since the early 1900s, music has always been innovated in a variety of aspects, no matter in style, chords, and instrumentation. We can apply data visualization on audio features and years to figure out how music has changed every year and do some research about these changes based on the development of pop music and EDMs in the 20th century and the 21st century.
- **What are the common features of the k-pop genre? and how is it different from average songs?**
    - Different song genres have their own characteristics. These characteristics can be determined by audio features of songs. We are interested in seeing whether

these features have correlation with the genre. In this project we want to primarily focus on the k-pop genre as it is a non-English music industry that has thrived globally over the past few years. We are interested to see if the features of K-pop songs, which keep growing, correspond with the overall music trend over time or not.
    - **Result:** K-pop songs follow the key attributes of popular songs including higher danceability, energy, and valence, and lower acousticness.

- **Do song features account for the difference of the popularity between k-pop boy-group and girl-group songs?**
    - It has been commonly known in the K-pop community that boy bands usually yield higher revenue than girl-groups in general. We want to know if audio features are one of the factors that causes the popularity difference and if people do spend more time listening to boy-group songs not just because of its album sales. We can compare the mean of each feature in boy-group and girl-group songs with the average features of all songs by year to see if there is some correlation.
    - **Result:** K-pop boy-group and girl-group songs are not any more popular than other k-pop subgenres. Although boy-group and girl-group songs follow the key attributes of popular songs, users do not spend time listening to their music any greater than other k-pop subgenres.

Datasets:

- 1st dataset: Audio features of 160,000+ tracks on Spotify
  https://www.kaggle.com/yamaerenay/spotify-dataset-19212020-160k-tracks
- 2nd dataset: The top songs by year in the world by Spotify.
  https://www.kaggle.com/leonardopena/top-spotify-songs-from-20102019-by-year
- 3rd dataset: K-pop music videos.
  https://dbK-pop.com/db/k-pop-music-videos

Challenge Goals:

In our data science project, we are looking to meet these following challenges

- **New libraries:**
    - Our research questions require a lot of data visualization. It would be more helpful if we manipulate and customize data visualizations more widely. Thus, we are going to use a new library called Altair for our data visualization. Moreover, we are also going to retrieve data through Spotify's API. More supplement libraries will be needed in order to access the web database as well.

- **Multiple datasets**
    - Our dataset with audio features is not completed. This dataset does not involve song genres. We will need to find supplement datasets to answer our research questions. We have found a dataset with K-pop songs up until the present. More datasets of other genres will be added as we work through to perform machine learning with a variety of genres.

Methodology:

- **Part 0: Get access to the data using Spotify API**
    - First we have to write a class that helps authenticate access to Spotify's web API. Later on, we are going to write several class functions that help retrieving different types of data such as album, track or artist data.
- **Part 1: Overall spotify songs analysis:**
    - Is there any commonality among popular Spotify songs in the year they were released? and has the trend changed over past years? (Dataset: 2nd dataset)
        - We are going to use the Altair library to create a scatter plot to represent the density of features and thus get the commonalities among the features of songs within a year.
        - We are also going to use the Altair package to plot a line to visualize the general trend of features over years. In this case we are using the y-axis as the mean value of features. (e.g. the mean BPM of a song)
    - How have typical features of songs changed over time? (Dataset: 1st dataset)
        - Since the dataset included a lot more songs than the 2nd dataset which only contains the top songs. We are just going to find out the general trend of the features and combine what we've found to the popular music history. We are going to do something very similar to the second part of the first question (i.e. using a line plot to represent the trend) but we are going to look at the result in a more generalized way.
- **Part 2: Spotify K-pop songs analysis:**
    - Generating a new dataset with only K-pop songs and their audio features.
        - Our 1st dataset which contains over 160,000 Spotify songs does not label song categories. Since we want to specifically get audio features of K-pop songs only, we are going to use another dataset (3rd dataset) that contains all the K-pop songs released. Next, we are going to go over every song's name in the dataset and search it on Spotify's web API to retrieve

its audio features. Once we searched, we are going to take the first song that appears assuming it is the closest matching result. However, searching does not always return a valid result. We need to further get the id of that song's artist and see if the artist has "K-pop" labeled or not. This is to reassure that all the songs we have gotten are actually K-pop songs. Once we have obtained all the K-pop songs we need, we can now convert into a new dataframe.

- What are the common features of the k-pop genre? and how is it different from other genres?
    - We are going to make a line plot corresponding to each different feature of kpop songs with years as the x-axis and compare it to that of all Spotify songs.
- Do song features account for the difference of the popularity between k-pop boy-group and girl-group songs?
    - For this question, we are going to make a line plot with multiple series corresponding to girl-groups, boy-groups and others. Each plot will represent each feature. We are going to use these plots to see if the commonality within boy-group songs resembles the common audio features of top songs on Spotify as well. In addition, we will also want to see if the boy-group songs are actually more popular than girl-group songs.

Results:

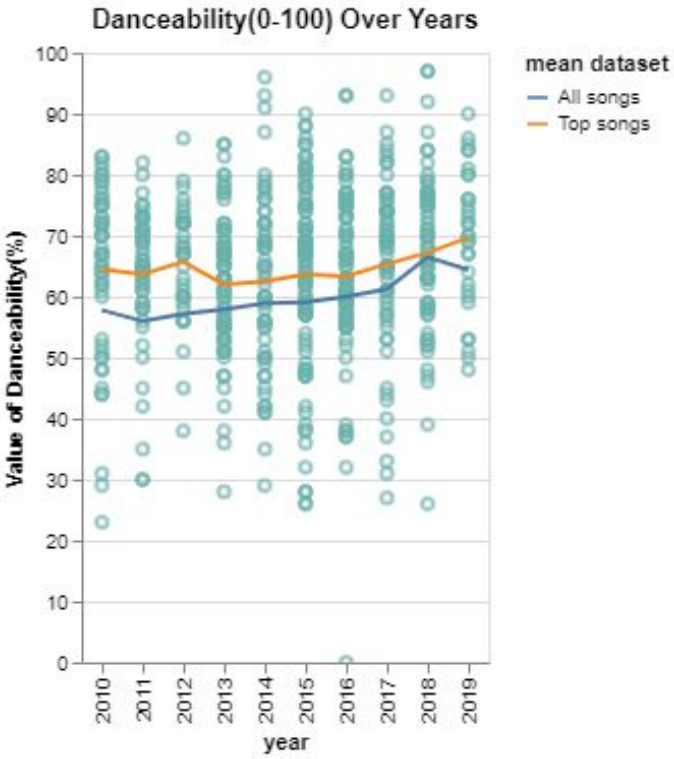
As a background, our datasets contain multiple audio features with the definitions from Spotify Web API Reference as followed. (source: https://developer.spotify.com/documentation/web-api/reference/tracks/get-several-audio-features/)

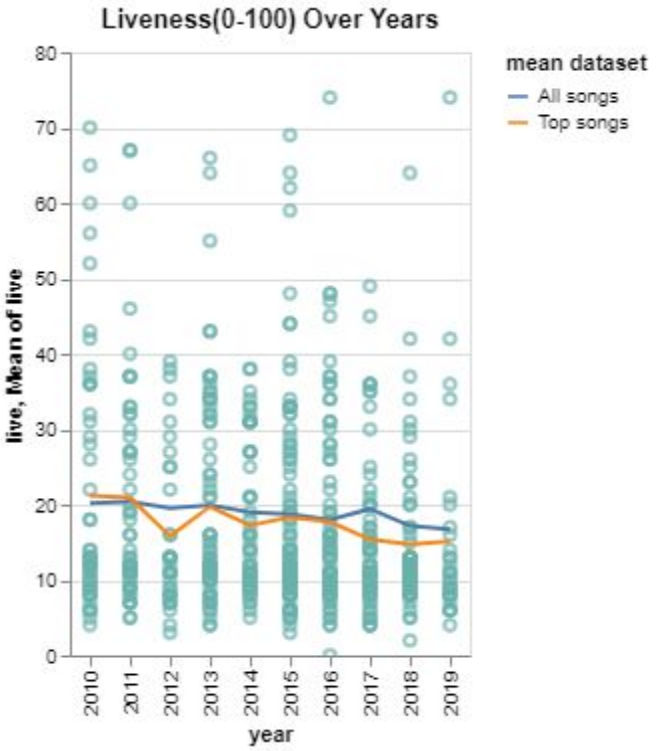
- **Acousticness** A confidence measure from 0.0 to 1.0 of whether the track is acoustic. 1.0 represents high confidence the track is acoustic.
- **Danceability** Danceability describes how suitable a track is for dancing based on a combination of musical elements including tempo, rhythm stability, beat strength, and overall regularity. A value of 0.0 is least danceable and 1.0 is most danceable.
- **Energy** 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.
- **Liveness** 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.
- **Speechiness** 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.
- **Valence** 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).

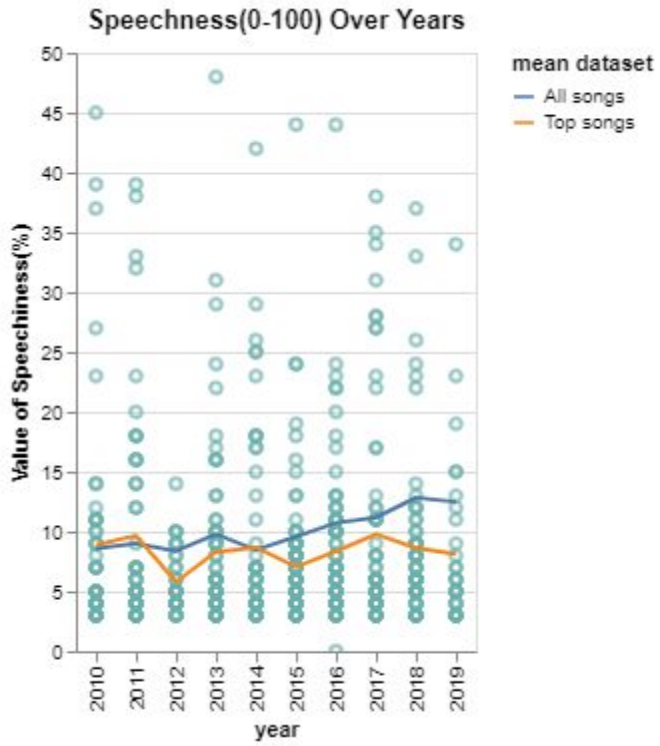
Part 1: Top spotify songs analysis

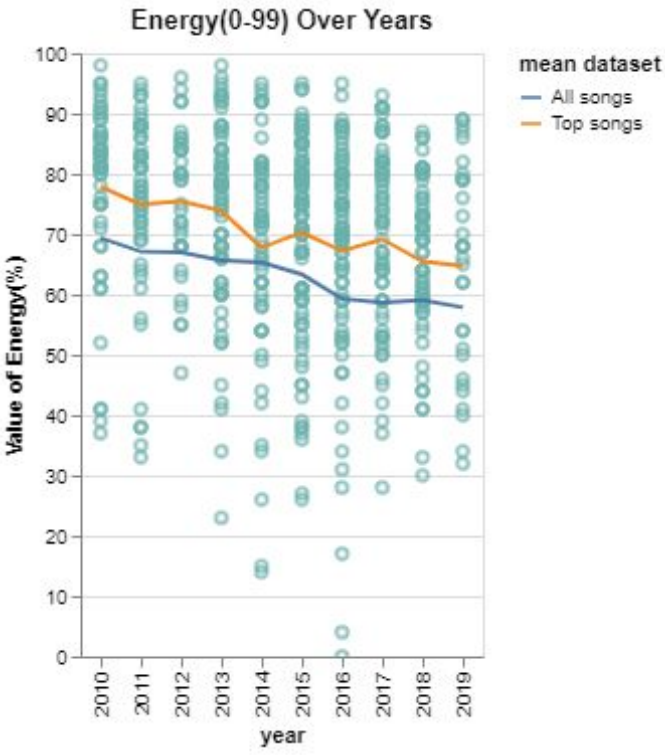
For part one, we used a scale from 0 to 100 to make the consistency between two datasets.

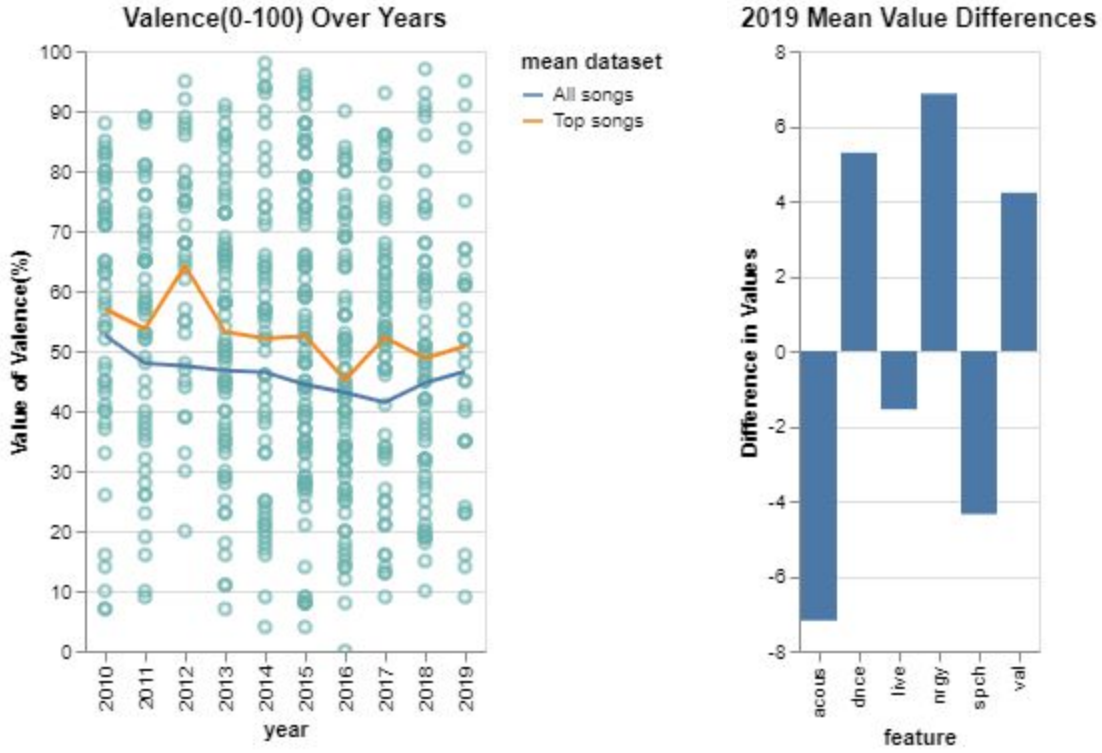
- **Part 1.1**







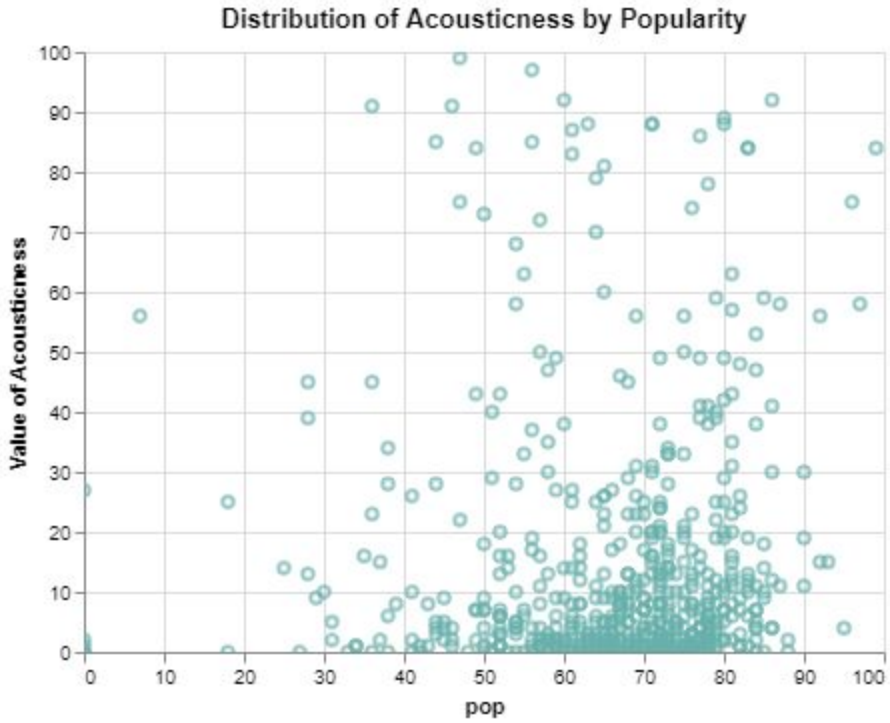


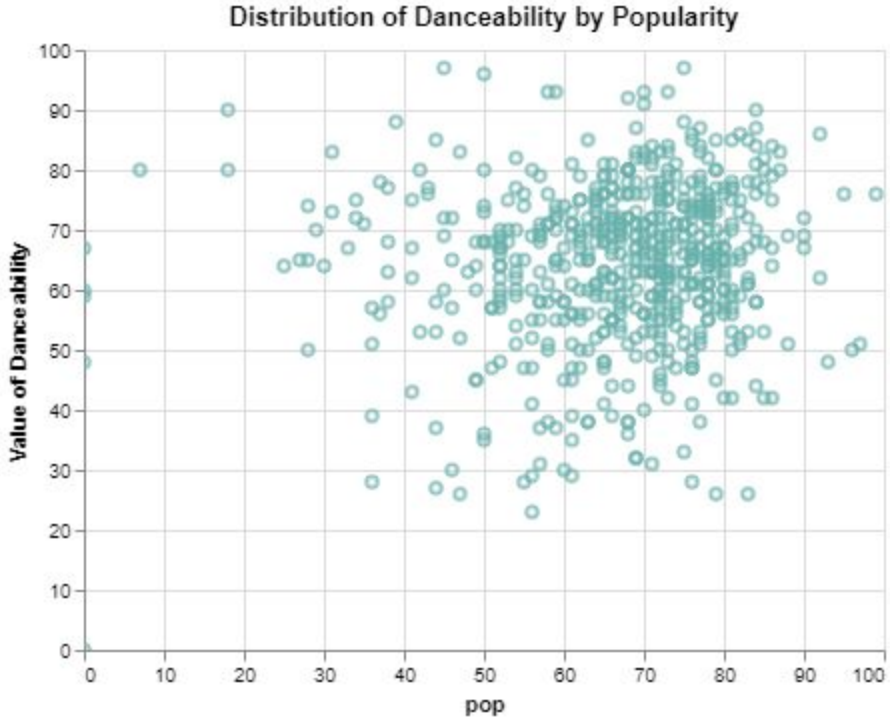


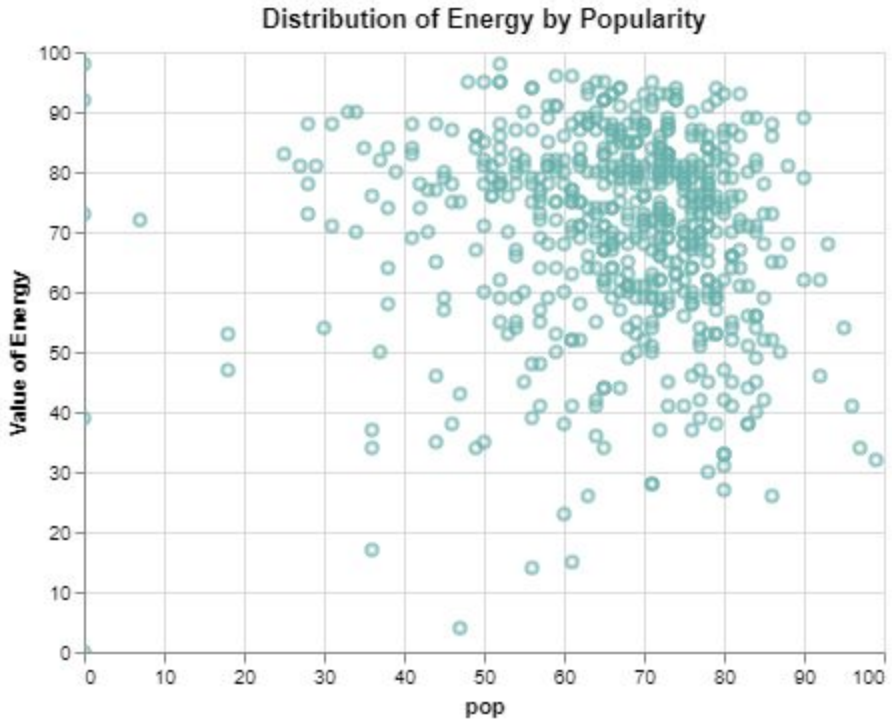
- The first set(with green points and lines) focused on the general trends of the top songs compared to the total 160k songs from 2010 to 2019. The green points showed the distribution of the values of the features. We took six features of songs to analyze. The trend of all plots looks rather obvious: from 2010 to 2019 the mean values and density of features do not vary too much(i.e. No sudden increase/decrease). The reason for that might be that the time interval is too small from 2010 to 2019, so we would have two additional charts to analyze in part 1.2 that focus on the mean values of two features from 1921 to 2020.
- Another thing to be worthy of notice is about the difference between the mean value of features of top songs and that of all the 160k songs. The mean values of features of tops songs seem to follow the trend of that of the 160k songs. But their values differ in a certain distance. So we created an additional chart showing the differences of mean value of features between two datasets in 2019. Along with the 6 graphs of features over years, we found that:
    - Top songs tend to have more energy than the general songs.
    - They are also likely to be more electronic(lower acousticness).
    - The moods of the songs are happier(higher valence).

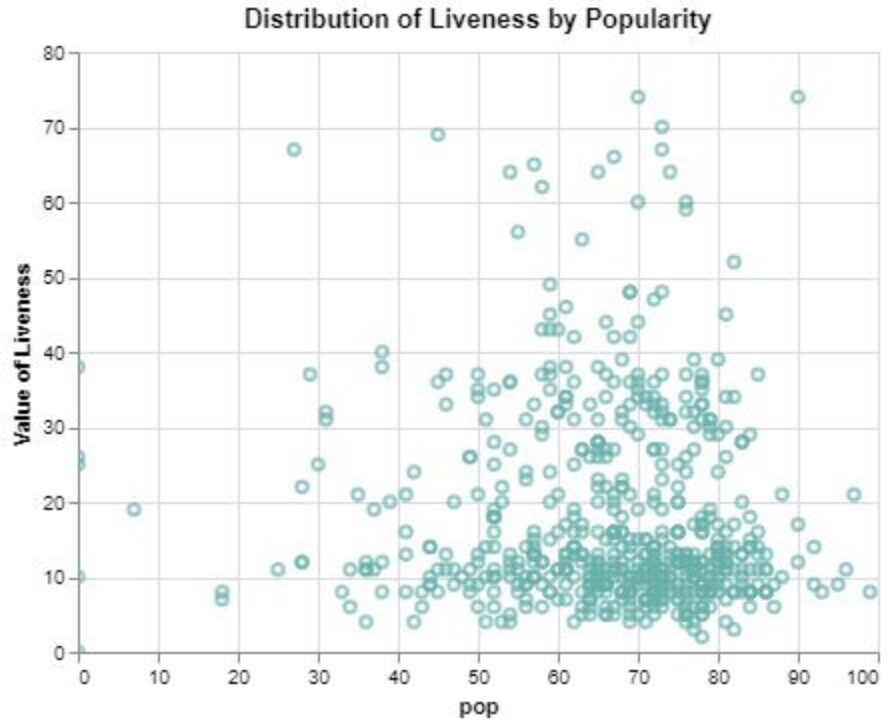
To show the values of features that have the greatest popularity, we conducted Part 1.2 focusing on the distribution of values of features among popularity.

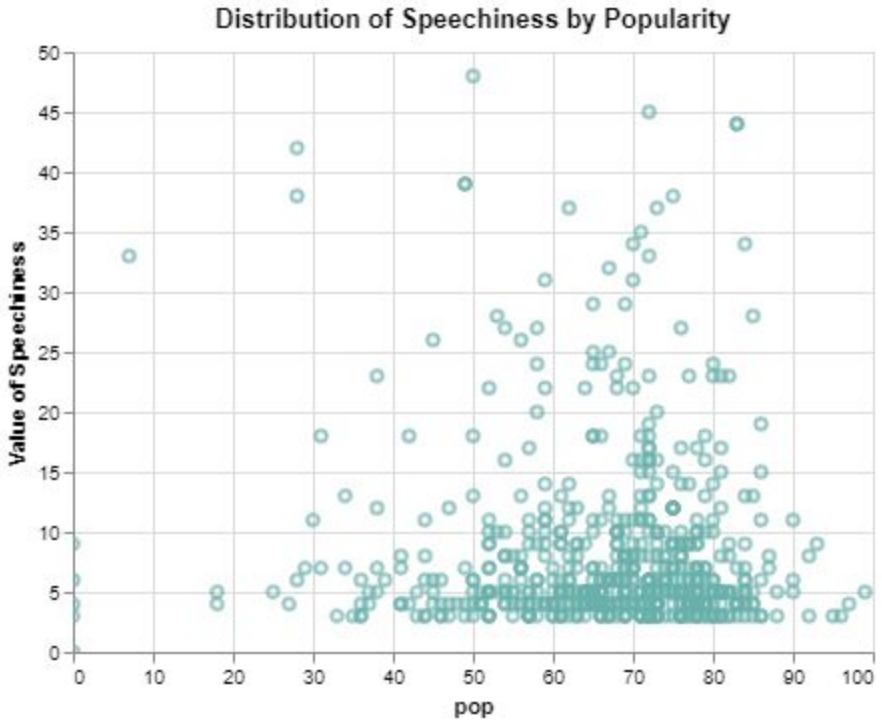
Part 1.2

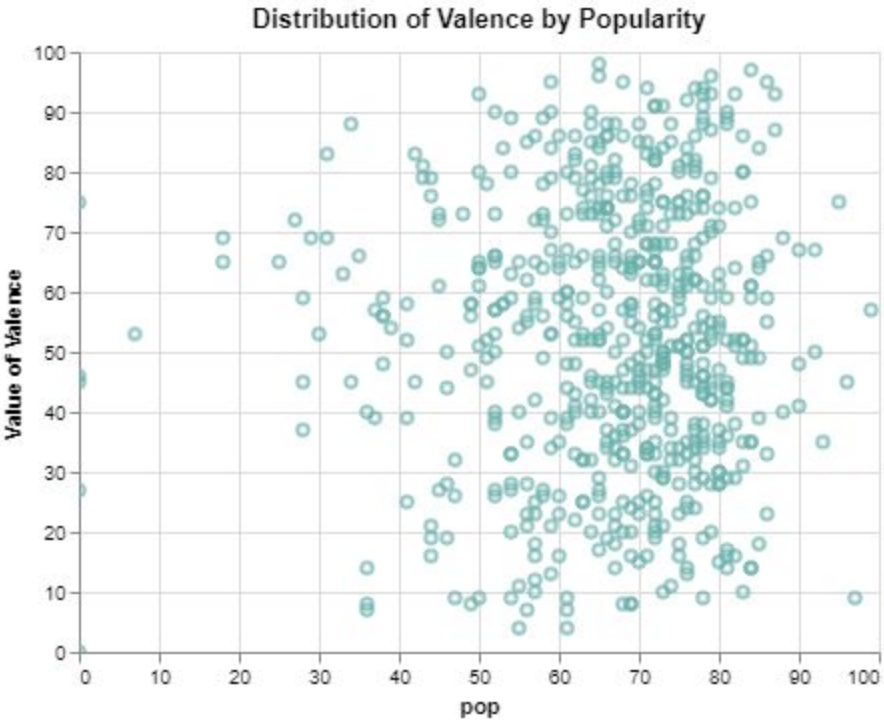

Distribution of Acousticness by Popularity
Value of Acousticness
pop
0
10
20
30
40
50
60
70
80
90
100

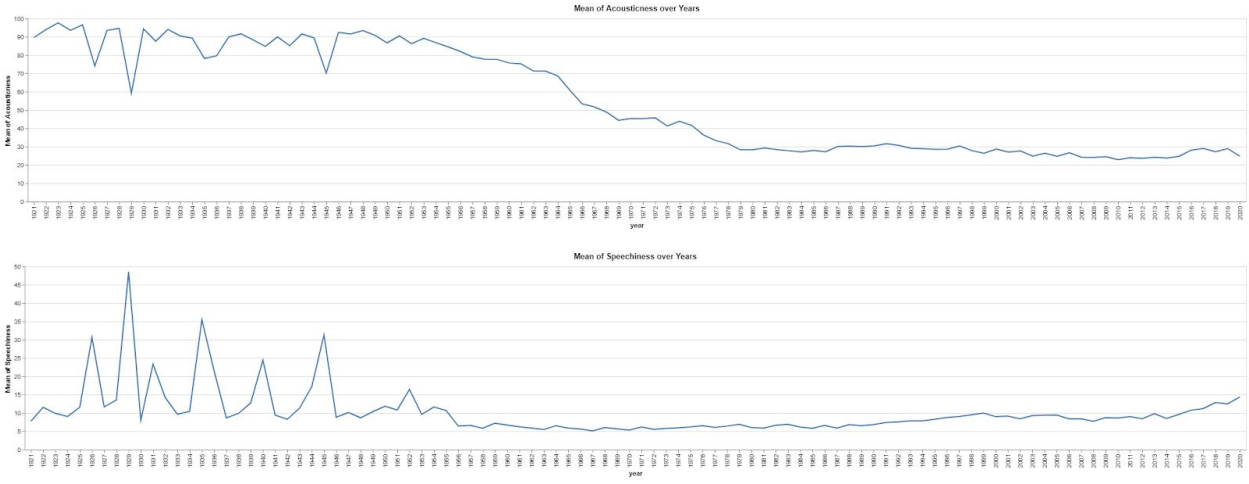

Distribution of Danceability by Popularity
Value of Danceability
pop
0
10
20
30
40
50
60
70
80
90
100


Distribution of Energy by Popularity
Value of Energy
pop
0
10
20
30
40
50
60
70
80
90
100


Distribution of Liveness by Popularity
Value of Liveness
pop
0
10
20
30
40
50
60
70
80
90
100


Distribution of Speechiness by Popularity
Value of Speechiness
pop
0
5
10
15
20
25
30
35
40
45
50
0
10
20
30
40
50
60
70
80
90
100


Distribution of Valence by Popularity
Value of Valence
pop
0
10
20
30
40
50
60
70
80
90
100
0
10
20
30
40
50
60
70
80
90
100

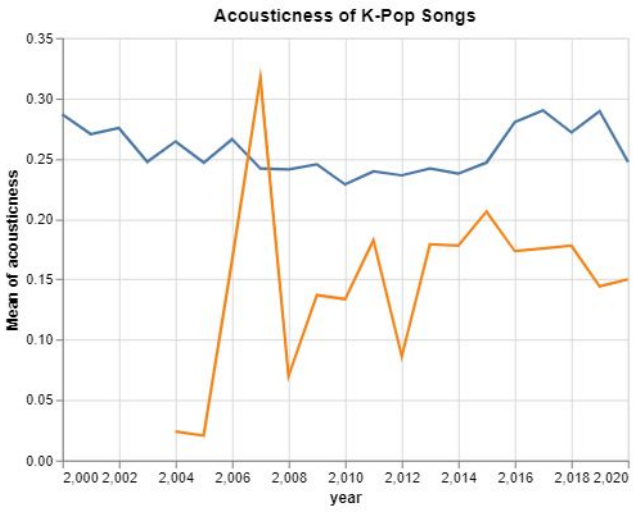


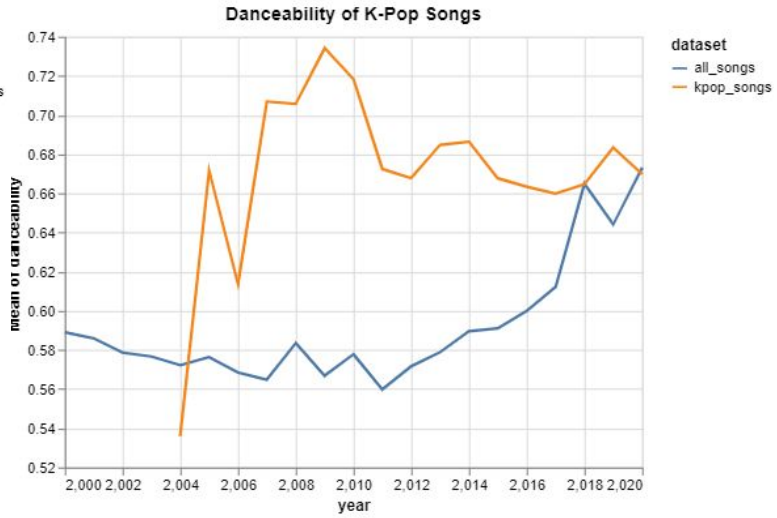
- In Part 1.2, we focused on how people's favourite songs' corresponding feature values and created 6 scatter plots of feature values among popularity in the top songs dataset. We got some results that actually made a lot of sense:
    - The distribution of acousticness of songs is mainly focused on low values, since most songs recently are more electronic and include a lot of synthesized instruments and electric instruments. And one interesting thing is that most of the points were concentrated around 0, which really showed that electronic music has really taken the mainstream of the music markets. But some points with high popularity values also have high acousticness, this really showed that the popularity among the acousticness is really polarized and there are people who still love classical acoustic music.
    - The points of danceability of songs are concentrated quite high in the second graph. It looks like most people tend to love rhythm/groove centered music since the rise of social dance from the 1930s. They treated music as a way to socialize and entertain.
    - Most of the energy points concentrated at a high level. This makes sense since by the two factors we mentioned above, the danceability and acousticness, do affect the energy of a song. Because if we want to have high danceability we need to create grooves, by that we need high energetic kicks and hits. So no doubt the average of energy is relatively high.
    - Liveness is a little tricky if we look at it, since most of the points are concentrated on the lower level and the highest value of liveness is still below 80. We searched about the liveness of a song and found that a value of liveness more than 80 is considered 100% to be a live song. But since we are analyzing the top songs of each year, it's not likely to have high liveness songs.
    - The distribution of speechiness is focused on low levels. We will focus on the speechiness more specifically later, but for here it is true that most of the songs do not contain too many spoken words.

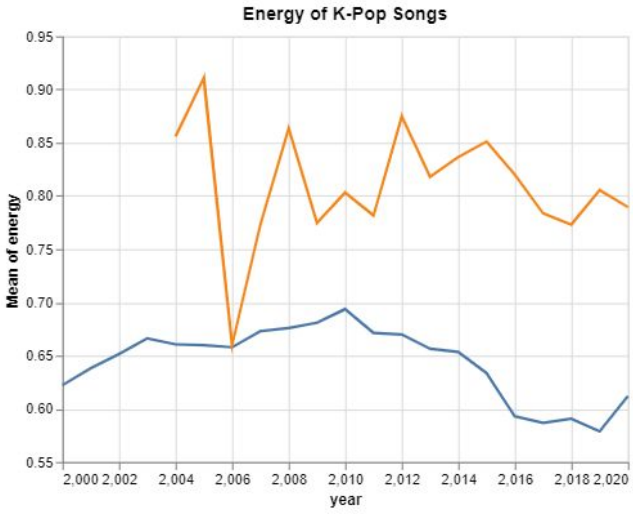
    - Valence is an interesting one, among all the one sided distributions, this one is quite balanced. Valence means the mood of a song, with a higher valence the song is more positive in mood and with a lower valence the song is more negative.
- We also focused on two specific features that made me feel confused at first, so we created two line plots to show how the mean of the values of these two features varied from 1921 to 2020 in the 160k dataset.
    - The first graph showed the mean values of acousticness over 100 years. We can see that from around 1953 the mean values of acousticness began to decrease until 1979, then the mean of acousticness kept at a relatively low value. Since one of us has learned class MUSIC 163 which studied the history of popular music in America. We think this graph fits the history of pop music. During 1940-1950, American music was popularized by Rock & Roll music and Rhythm & Blues music. More electrical instruments were introduced at that time. In around 1970s Disco music and synthesizers were introduced, thus more and more music started to have these electronic sounds as instruments and this brought the average acousticness to a lower level of presence in songs after the 1970s.
    - We were confused about the speechiness of the song at first, we thought that it would be a value to detect the lyrics that have no music pitches, like the rap songs. But it tends to be the value of spoken words in a song, so we have made a graph showing the mean speechiness over 100 years. This one is quite interesting since the mean values were almost stable except the year from 1925 to 1960. Since high speechiness songs are probably just talk shows and radio shows, combining with the insight from MUSIC 163, radio was popular around the 1920s while TVs were popular around the 1950s. This really explained the ups and downs during those years.

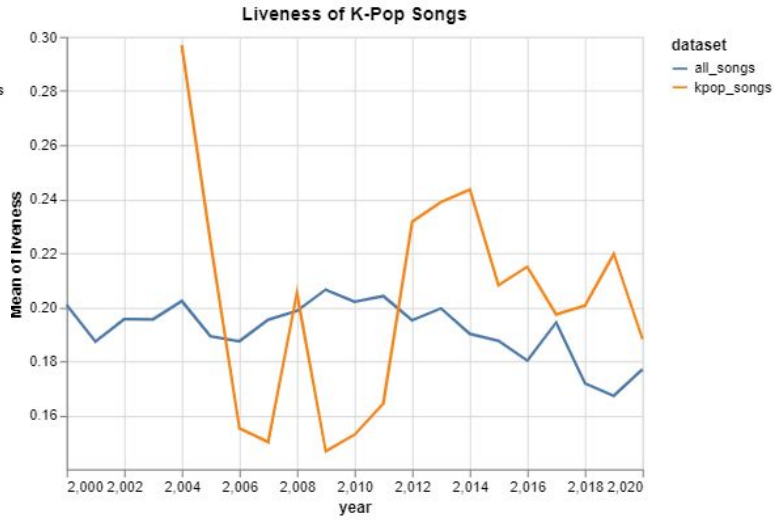
Part 2: K-pop songs analysis

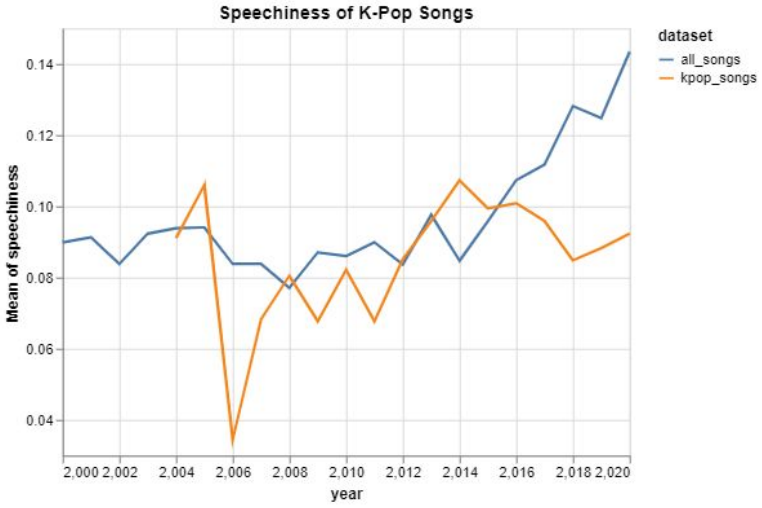
Part 2.1: Comparison between k-pop songs and average songs

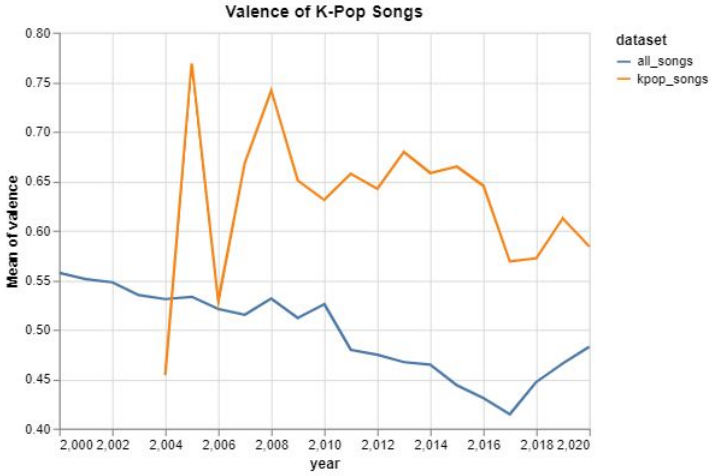












- From the visualization, we can see that k-pop songs tend to have less acousticness and speechiness while having more danceability, energy and valence than the average songs. Kpop songs previously tended to have less liveness and got more than the average songs later on. However, the line is fluctuated with a possible cause from relative lower data size compared to all the songs. From this analysis, we can see that k-pop audio features follow key attributes of popular songs from the previous part which includes higher energy, danceability and valence, and lower acousticness. This implies the popularity this genre gained in the past years still rely on these features.

Part 2.2: Comparison between k-pop sub-genres

- Although boy-groups and girl-groups are the backbone of the k-pop industry that contributes larger revenue including album sales and merchandising, we surprisingly found that the popularity of these two sub-genres on Spotify is overall slightly lower than the other k-pop songs including co-ed and solo songs. However the lines are very fluctuated that we can hardly deduct any trends from. K-pop boy-group songs have slightly higher popularity like we expected from its profitability. However the audio features and popularity are not really associated like from part 1. First of all, we can see that boy-group and girl-groups songs follow almost all the key attributes of popular songs in part 1 and significantly higher than any other k-pop sub-genres. However, their popularity does not differ at all. Another thing to take away is that although girl-groups songs follow the key attributes of popular songs better than boy-group songs including higher energy, danceability and valence, it is difficult to say with these visualizations that these features contribute to their popularity.
- In conclusion, the k-pop genre may depend on many other factors. The profitability of boy-groups and girl-groups does not mean fans will listen to their songs more on a streaming platform. The k-pop industry undeniably relies on many other goods beyond songs. Visuals, performance, and stories etc. are many other factors that contribute to this industry.

Work plan Evaluation:

- **Communication and File sharing**
    - We will be communicating through the Discord voice chat regularly and proceed with our project.
    - We are going to use Google Drive to store and share our datasets and documents, and we are going to share data analysis code using GitHub in Jupyter notebook format.
- **Member responsibilities**
    - Since we have two members, our tasks will be mainly divided by Part 1 and Part 2 of the project.
    - Phoenix Yi will be doing Part 1 which is focused on overall songs on Spotify. He will also be responsible for reading Altair's documentation and sharing it to another member.
    - Laphon Premcharoen will be focused on K-pop songs starting from getting K-pop song's data through Spotify's web API and working on Part 2 afterward.
- **Tasks:**
    - Finished by August 9th
        - Writing a class to access Spotify's web API for Part 0 (4-5 hours)
            - Since web API is a new concept for us, we first need to study both the endpoints of Spotify's web API and also how to write a class to access it.
        - Learn Altair library (2-3 hours)
    - Finished by August 18th
        - Data visualization for Part 1 (5-6 hours)
        - Data visualization for Part 2 (5-6 hours)
        - Analyze and report (2-3 hours)
            - We are going to put all the necessary charts into one document and analyze them. Finally, we can write a report for our findings.
- **Evaluation**
    - We were able to follow the deadline for the work we planned correctly. Each member worked on the plans as expected.
    - However, we did not set a plan for testing and that our Jupyter notebooks needed to be converted to .py files at the end. Out schedule was tight on this one since, these tass took a significant amount of time to finish

Testing:

- **Part 1**
    - Since the whole part 1 used Altair and Pandas to process datasets and charts, it is a little difficult to test all the functions using asserts. So we selected a few rows from the original datasets and made our testings based on them.
    - Our testing file will generate several testing charts that show if functions we wrote in Part 1 work well. We sliced 500 rows from the 1st dataset and 200 rows from the 2nd dataset and they worked pretty well and generated some readable plots.
    - For the second function the chart did not seem to be readable since we set the wrong parameters and caused a lot of inefficiency. So, we revised the algorithms and got a faster and more generally-useful function.
- **Part 2**
    - For this part we have done similarly to Part 1, we cut the 1st dataset and the k-pop dataset obtained from API into 5 rows each. Each row will be a representative from each year. By putting on tooltips on the visualization we could figure out if the lines correspond to the datasets correctly.

Collaboration:

- Since we have only two members, it is easy for us to coordinate and share responsibility. We did not have to have a leader in doing so.
- Discord turned out to be a great social media platform to collaborate with its flexible voice calls and messaging.
- We studied the fundamentals of Spotify's API from a YouTube channel. Afterward, we looked up the documentation ourselves on Spotify's website to write functions that meet our needs in this project.
- Many other programming forums such as StackOverflow were helpful to us as well. These websites helped us go through any issues we had when writing the codes.
- For the terms of features in the Spotify songs, we used an online source explaining each term's meaning and how does it matter. We will attach the link to that web page in the Addition Source page.

Addition Source:

- https://en.wikipedia.org/wiki/Spotify
- https://developer.spotify.com/documentation/web-api/reference/
- https://towardsdatascience.com/is-my-spotify-music-boring-an-analysis-involving-music-data-and-machine-learning-47550ae931de#:~:text=Speechiness%3A%20%E2%80%9CSpeechiness%20detects%20the%20presence,does%20not%20have%20any%20speech.