Spotify: What Makes a Top Song?

DigHum 100 Dr. Anderson | Gloria Yang | June 10, 2021

Project Description: Modernization has led to different approaches and greater diversity of sound characteristics and range. In this project, I will be analyzing popular songs on Spotify over time and looking for common attributes and trends. If we can achieve insights from the data as to what attributes and trends are helpful for predicting a song's popularity, this will help musicians to understand what listeners prefer and how to compete in the music industry.

Dataset:

"Spotify: All Time Top 2000 Mega Dataset" is a dataset taken from Kaggle which contains audio statistics and attributes for Spotify's top songs. The dataset is contained within a csv file and includes data on songs released from 1959 to 2019. The data set breaks down each song by the following attributes: genre, year, BPM, danceability, energy, loudness, liveness, valence, length, acousticness, speechiness, and popularity.

Title: Name of the track
Artist: Name of the artist
Top Genre: Genre of the track
Year: Release Year of the track
BPM (Beats Per minute): the tempo of the song
Energy: the higher the value, the more energetic the song
Danceability: the higher the value, the easier it is to dance to the song
Loudness (dB): the higher the value, the louder the song
Liveness: the higher the value, the more likely the song is a live recording
Valence: the higher the value, the more positive the song
Length: the duration of the track
Accousticness: the higher the value, the more acoustic the song
Speechiness: the higher the value, the more spoken word the song contains
Popularity: the higher the value, the more popular the song is

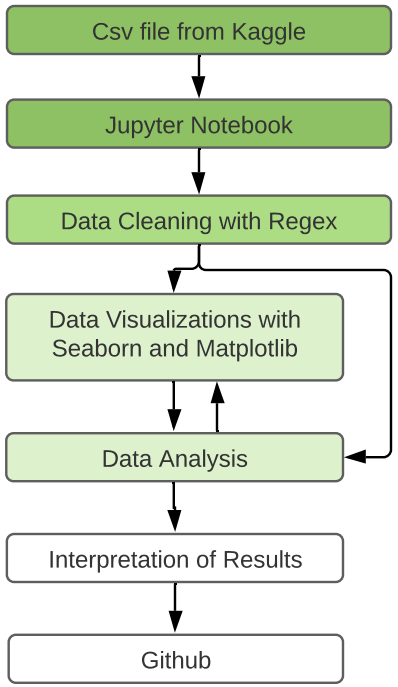
Index	Title	Artist	Top Genre	Year	Beats Per Minute (BPM)	Energy	Danceability	Loudness (dB)	Liveness	Valence	Length (Duration)	Acousticness	Speechiness	Popularity
1	Sunrise	Norah Jones	adult standards	2004	157	30	53	-14	11	68	201	94	3	71
2	Black Night	Deep Purple	album rock	2000	135	79	50	-11	17	81	207	17	7	39
3	Clint Eastwood	Gorillaz	alternative hip hop	2001	168	69	66	-9	7	52	341	2	17	69

Questions for EDA:

- Can we predict a song's popularity based on attributes such as danceability, valence, loudness, and length?
- Are there any trends in the popularity of different music genres?
- How do song attributes differ according to different genres?
- How have the attributes of songs on Spotify's Top 2000 changed over time?

Tools:

- **Matplotlib and Seaborn:** libraries for generating visualizations for data analysis
- **Pandas:** library for storing, cleaning, and manipulating data within dataframes
- **Jupyter Notebooks:** tool for data cleaning, data manipulation, and developing visualizations
- **Regex:** implemented to clean genre data from subgenres
- **Github:** portfolio repository that includes the dataset, storyboards, Jupyter notebooks, etc.



Methodology:

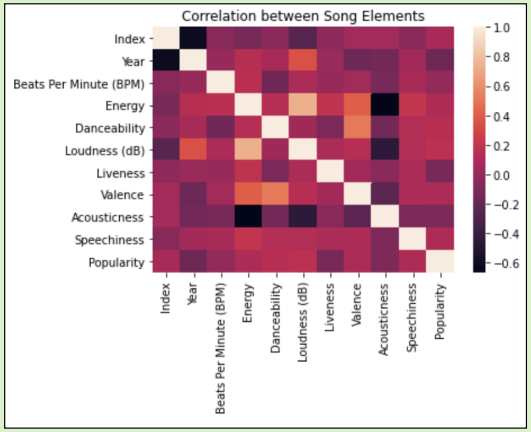
1) I will be converting the "Spotify: All Time Top 2000 Mega Dataset" .csv file into dataframe structure. Tools such as Matplotlib and Regex will also be imported.

2) Using regex, songs will be categorized from 10 main genres; songs that have other genres will be categorized as "Other". Data will be further cleaned and formatted.

3) I will be exploring correlation and typical ranges for different song attributes. A heatmap will first be derived from the dataframe to see if there are any correlations between any two song attributes.

4) Using Pandas and data visualization, I will explore shifts in song attributes and genre distribution over time. If there are existing shifts or common attribute characteristics among popular songs, I will try to see if and which attributes are helpful in predicting a song's popularity.

Results:



I first created a heatmap to get a beginning idea of existing correlations between song attributes.

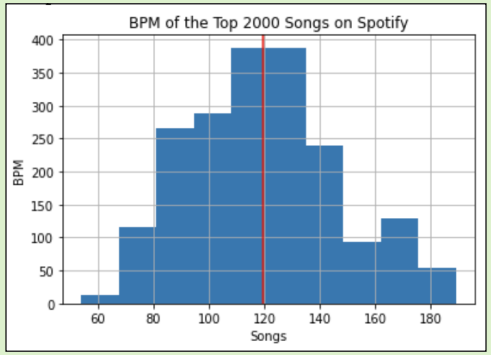
A couple of insights we can derive from the above heatmap:

1) My heatmap indicated a strong correlation between loudness and eneregy. Analysis on a scatter plot showed the correlation was positive; the more energetic a song, the more likely the song is to feel loud.

2) Acousticness has a negative correlation with energy and loudness.

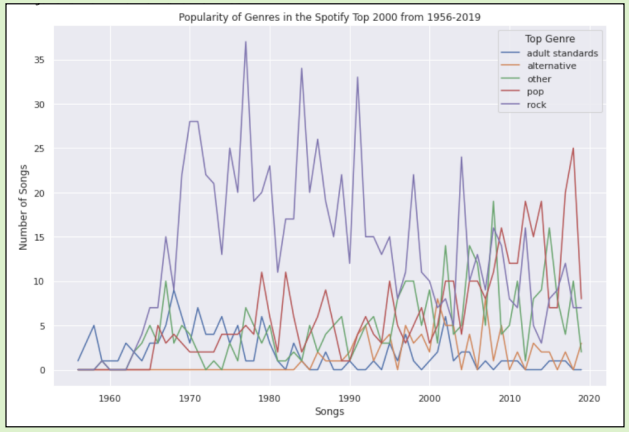
3) Valence and danceability are highly correlated. This makes sense, as dance songs are often happier and in a major key.

I was also interested in seeing whether popular songs had ideal lengths and BPMs. Using histograms, I found that the mean song length was approximately 4.15 minutes but that there was a lot variation possible due to different standards for different genres. After removing outliers, I also learned the majority of songs had a BPM within a range of 100-140.



I then used regex to categorize subgenres into the following genres: adult standards, alternative, country, electronic, folk, funk, hip pop, indie, metal, pop, rock, soul, and other. Across all songs in the dataset, the most popular genres were different types of rock and pop.

Results Cont:

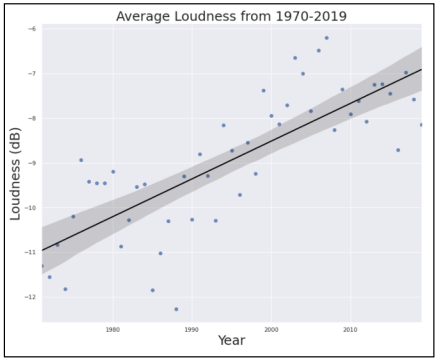


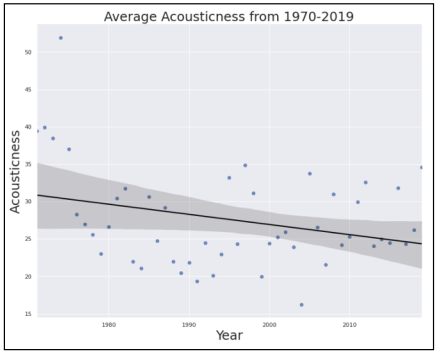
However, in terms of popularity prediction, attributes in the dataset are taken from top songs over a broad period of time and might not be representative of songs in the present day. Modernization has led to different approaches and greater diversity of sounds such as louder sounds, faster tempos, and speechiness. We must therefore also examine trends in changes to attributes over time.

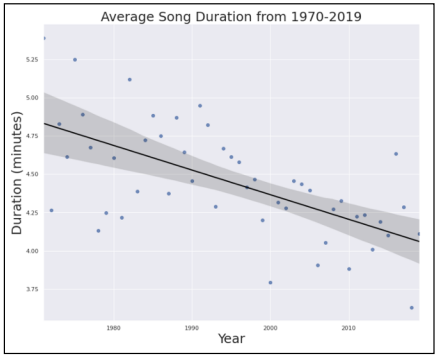
The graph above shows the number of songs included in Spotify's Top 2000 over time in the following genres: adult standards, pop, rock, alternative, and other. While the number of of rock songs decline into the 21st century, genre diversification is shown through growth for other genres in the number of songs in Spotify's Top 2000 over time. Due to this genre diversification, I expected to see additional shifts in song attributes such as length and BPM.

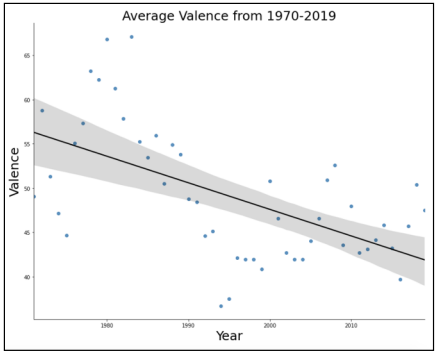
Over time, we can see that the most popular genre shifted from rock to pop. When I examined other attributes for trends over time, I also noticed that attributes such as accousticness, valence, and duration tended to decrease while attributes like loudness and da tended to increase. According to The Atlantic, music, particularly pop music, has become intrinsically louder due to current trends in the industry; valence has also decreased due to widespread content for more negative and sadder content. Likewise, due to new technology, electronic sounds have become more prominent over time. Music has become more electronic and less acoustic, mirroring our modern society's integration with technology. With further research, I also learned that due to the emergence of streaming platforms like Spotify, artists are paid by how many songs are streamed rather than physical sales. This then leads to shorter songs and new song structures such as pop overture. Therefore, besides gender diversification, a song's popularity is also influenced by emotional preference and technological and streaming modernization.

Graphs for song attribute trends over time can be viewed in the right column.





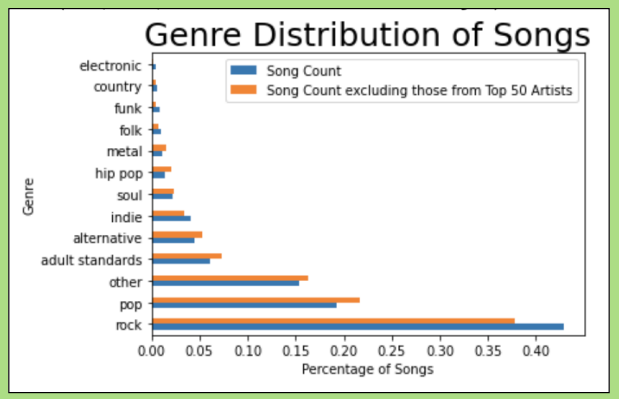


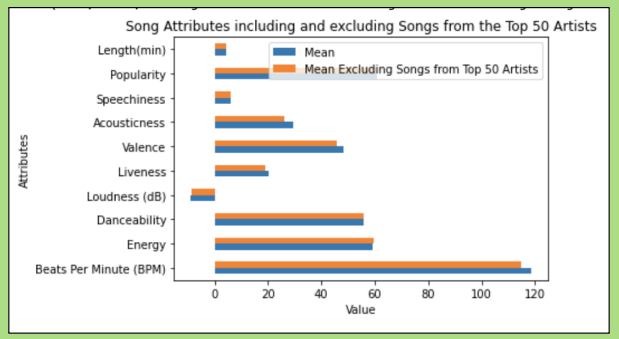


Addressing Potential Holes:

There was a possibility that the my observations were due to the popularity of artists who had many songs included in the dataset rather than because of overall music shifts. If there were too many popular artists who had a large number of songs listed, this could skew the dataset's overall genre distribution and other attributes due to less diverse production. In order to see if popular artists had skewed the distribution, I compared attribute and genre distribution of the overall dataset with and without the dataset's top 50 artists. 702 out of 2000 songs were from the top 50 artists and the majority of artists produced rock songs.

Genre distribution stayed relatively the same when excluding songs from the top 50 artists with a slightly lower percentage of rock songs due to exclusion of rock artists like Rolling Stones, Queen, and Bruce Springsteen.





However, if we compare song attributes when excluding and including songs from the top 50 artists, we can see that there is little difference between the song attribute means. Therefore, the inclusion of artists with multiple songs did not skew our observations regarding attribute and genre distribution and trends over time.

Conclusion:

If we were to predict the popularity of a song based on its attributes without the context of time, we would be able to make an better approximation based on whether it had rock or pop as a genre. The genre would have also influenced other attributes that we could use to gauge the song's popularity such as whether the BPM was within 100-140 or whether it had a length around 4.15 minutes.

However, we've learned through our analysis that there has been an increase in genre diversification into the 21st century. The presence of genre diversification has in turn led to more variation for typical ranges in song attributions, particularly danceability, length, loudness, and energy. There's also been a shift of the most popular musical genre from rock to pop from the mid 20ths century to today. Therefore, if we were to predict the popularity of a song released in the modern day, the song would likely be more popular if it were a pop song as opposed to rock.

Although there is a possibility that we there are reccuring artists that may have skewed genre and attribute, we compared the dataset when including and excluding these artist and found minimal difference in genre distribution and attribute means for different genres. We can therefore conclude that our dataset when including songs from the top 50 artists was representative of all 2000 songs in the dataset.
Different genres have different means and typical ranges. As genres of popular songs have become more diversified, this has led to a shift in the mean of attributes. For example, we learned that danceability and loudness has increased over time while song duration has decreased over time.

Limitations:

Spotify is also a platform that was released 2008, but songs tend to receive more listens at the beginning of its release. Older songs that are on the platform but that were released before the platform existed would receive less listens than if the platform had released at the time of song's release. This would have made older songs potentially less likely to appear in Spotify's Top 2000 songs.

According to Spotify's financial press release in 2019, Spotify had 217 million active users. However, in comparison to other streaming platforms as shown on the right, Spotify had differing demographics in terms of age, gender, and country or origin. Because different demographics might listen to different times and amounts of music, my observations are not representative of all music listeners. There are also other ways for listening to music besides streaming platforms such as radio or buying albums. In order to better guage a song's popularity and how popular songs have changed over time, I would thus also have to include data taken from other streaming platforms and music services around the world.

Demographic	Spotify	Apple	Pandora
18-24	26	17	11
25-34	29	23	28
35-44	16	22	21
45-54	11	15	17
55+	19	23	22
Male	44	56	41
Female	56	44	59

Lastly, my dataset was only comprised of the top 2000 songs on Spotify. Although I was able to see trends over time, it would have been helpful for me to also compare the attributes of popular songs to songs outside of those in the top 2000. This way, I would be to see whether a trend in what was deemed popular was simply due to modernization or because of preference amongst music listeners.

Further Questions:

- How do the genre distributions and attributes change with different demographics such as gender and age?
- How representative are popular songs for trends in the music industry?
- Are there any specific artists or genres that are continuing to lead trends in song attributes?
- How have song attributes changed in the past year as a result of COVID 19?
  - I am particularly interested in investigating attributes like valence to see if artists' songs are mirroring listener's feelings of loneliness and sadness

In the future as I become more familiar with modeling, I am additionally interested in developing a machine learning to predict a song's popularity by analyzing the different metrics within the dataset. As mentioned in the project description, this will allow musicians to develop a better understanding of music listeners' preferences and trends will increase their probability of competing in the market.

Hyperlinks:

- **Hyperlink to Jupyter:** https://colab.research.google.com/drive/1DsZ4s_TyKT_TuBdc2z3GnCYQGoZT_HXB?usp=sharing
- **Hyperlink to CSV:** https://drive.google.com/file/d/16BOmSQKc0WRAMaaHU9TBXTGERMbiz2uI/view?usp=sharing
- **Hyperlink to Github:** https://github.com/gloriiayang/DIGHUM100
- **Hyperlink to Slides:** https://docs.google.com/presentation/d/18E8Pr2RZCReb6DAq_4Fu68MVCJHfwzVMHza6LwRh9o8/edit?usp=sharing

Works Cited

Biss, Madars. “Rhythm Tips for Identifying Music Genres by Ear.” Musical U, Easy Ear Training Ltd., 14 Feb. 2017, www.musical-u.com/learn/rhythm-tips-for-identifying-music-genres-by-ear/.

Esty, Thomas. “Trends Over Time - Music Popularity Data Analysis.”*Google Sites*, sites.google.com/site/musicpopularitydataanalysis/trends-over-time.

Fadelli, Ingrid. “Using Spotify Data to Predict What Songs Will Be Hits.”*Tech Xplore - Technology and Engineering News*, Tech Xplore, 9 Sept. 2019, techxplore.com/news/2019-09-spotify-songs.html.

Hinkes-Jones, Llewellyn. “The Real Reason Music's Gotten So Loud.”*The Atlantic*, Atlantic Media Company, 25 Nov. 2013, www.theatlantic.com/entertainment/archive/2013/11/the-real-reason-musics-gotten-so-loud/281707/.

Iqbal, Mansoor. “Spotify Revenue and Usage Statistics (2021).”*Business of Apps*, Business of Apps, 2 Apr. 2021, www.businessofapps.com/data/spotify-statistics/.

Przybyla, Matt. “Predicting Spotify Song Popularity.”*Medium*, Towards Data Science, 3 Feb. 2021, towardsdatascience.com/predicting-spotify-song-popularity-49d000f254c7.

“Shareholder-Letter-Q4-2019.”*Financials*, Spotify AB, 2 Feb. 2020, investors.spotify.com/financials/default.aspx.

Singh, Sumat.*Spotify - All Time Top 2000s Mega Dataset*, Kaggle, 4 Feb. 2020, www.kaggle.com/iamsumat/spotify-top-2000s-mega-dataset.

Stuart. “Pop Music Is Louder, Less Acoustic and More Energetic than in the 1950s.”*The Guardian*, Guardian News and Media, 25 Nov. 2013, www.theguardian.com/technology/2013/nov/25/pop-music-louder-less-acoustic.