Project Report

Spotify Songs and Billboard Rankings

1. Introduction

According to Billboard[1], over 100,000 tracks are uploaded to streaming platforms every day. But what makes a song stand out? Specifically, what makes a song popular and renowned enough to be listed on Billboard's Hot 100 Songs of the 2010s decade? Could it be a song's beats per minute, its danceability, or perhaps its energy level and shorter length? As avid music lovers who have listened to over 120,000 minutes of music this year combined, we are intrigued by the characteristics that make a song top-rated.

For this project, we plan to use Kaggle data containing 584 top (unique) songs from Spotify between 2010 and 2019 that come with different characteristics of that song, along with Billboard's Hot 100 songs from the same timeframe. We will join the two datasets and analyze the data to identify correlations between song features and their rankings.

2. Data

This project uses two primary sources of data: Kaggle's Spotify songs with features for each song and Billboard's Hot 100 list of the top 100 songs released between 2010-2019.

2.1 Spotify Songs and Features

We collected data from Kaggle[2], which included various features of songs such as the song name (*song_name*), artist, track genre (*top_genre*), year the song was top-rated (*year*), beats per minute (*bpm*), energy (*energy*), danceability (*dance*), loudness (*dB*), valence (*val*), duration (*duration*), acousticness (*acoustic*), speechness (*speech*), and popularity score (*pop*).

Kaggle's dataset consisted of highly rated songs from the 2010–2019 decade. We analyzed these features to identify potential similarities between highly ranked songs according to Billboard. No web scraping was necessary since this dataset was precompiled. During data cleaning, we removed 19 duplicate entries, to have a total of 584 songs for analysis, and renamed a few columns to better align with our analysis needs—for instance, "title" was renamed to "song_name." These cleaning steps were performed in the *billboard_raw.csv* file.

2.2 Billboard Rankings

Billboard's website[3] features a page listing the name, rank, and artist(s) of the top 100 songs of the 2010s decade.

We used web scraping to extract data from this page, including *song_name*, *artist*, and *rank*. This data is essential for merging the two dataframes on the *song_name* column. Billboard is a renowned music industry authority, known for its charts that rank songs, albums, and artists based on metrics such as

---

[1] https://www.billboard.com/pro/how-much-music-added-spotify-streaming-services-daily/

[2] https://www.kaggle.com/datasets/leonardopena/top-spotify-songs-from-20102019-by-year?resource=download

[3] https://www.billboard.com/charts/decade-end/hot-100/

sales, radio play, and streaming data. Its flagship charts, like the Hot 100 and Billboard 200, are globally recognized as key indicators of musical popularity and trends.

For data cleaning, we removed the last 13 rows from our dataframe, as the web scraping process captured more data than needed. Additionally, we created a dictionary based on our web scraping list names and transformed the dictionary into a dataframe.

Unfortunately, a few days before our project was due, this website started to require a $200 subscription to access. With permission from Professor Colbert, we had to instead create an excel file and manually input the 100 song names, artists, and ranks. This was a sad hiccup in our project, but found in our *00_scrapeproject.ipynb* file is our original scraping code. This code was fully functional before the subscription became a factor. But now we are using the *billboard_raw.csv* file as our 'scraped' data.

2.3 Combining Kaggle and Billboard

Since both datasets included the same song names, we thought it easiest to merge on that feature. We decided to use an inner join because we only wanted to analyze songs that were listed on the Hot 100 list as they were the best songs of the decade. Therefore, this merged data frame contains the **intersection** of the data and only contains data that appears in both datasets. Once merged, we changed all column names to match the column names in our data dictionary (*Table 1*). We did realize there were some duplicate songs when we merged and so we dropped those. Further, we thought it would be an interesting statistic to see if multiple artists had an impact on song rank (*section 3.2*). So, we made a new column titled "multiple_artists" which is a binary column that shows a 1 if the song has multiple artists (feature or duet) and a 0 if the song is a solo.

Next, we created a new column titled "energy_bin" to categorize each song's energy level into 'Low,' 'Medium,' or 'High.' This categorization will be a key part of our analysis. We will be able to tell what genres are likely to be in a specific energy bin (*section 3.4*). The dataset, now refined with these categories, was exported to a CSV file titled *final_dataframe.csv*, containing 46 songs ready for observation. A detailed description of each feature is provided in Table 1, offering additional context for the analysis.

Table 1 Data Dictionary

Column	Type	Source	Description
song_name	Text	both	Title of the song
artist	Text	both	The creator of the song and any featuring artists
top_genre	Text	Kaggle	Genre of the track
year	Numeric	Kaggle	Year song was top rated
bpm	Numeric	Kaggle	Beats per minute, tempo of the song

energy	Numeric	Kaggle	The energy of song, the higher the value, the more energetic the song
dance	Numeric	Kaggle	Danceability, the higher the value, the easier it is to dance to this song
dB	Numeric	Kaggle	Loudness in decibels, the higher the value, the louder the song
val	Numeric	Kaggle	Valance, the higher the value, the more positive mood for this song
duration	Numeric	Kaggle	The duration of the song in seconds
acoustic	Numeric	Kaggle	Acousticness, the higher the value, the more acoustic the song is
speech	Numeric	Kaggle	Speechness, the higher the value, the more spoken words the song contains
pop	Numeric	Kaggle	Popularity, the higher the value the more popular the song is by users
rank	Numeric	Billboard	The ranking position on the Billboard top 100 chart between 2010-2019
multiple_artists	Binary	Both	A 1 in shown if a song has multiple artists featured in the song and 0 if the song is a solo
energy_bins	Categorical	Both	Labels the song as ‘Low’ energy if the song has an energy score between 0-49. ‘Medium’ if a song has an energy score between 50-75 and ‘High’ if a song has an energy score between 76-100.

3. Analysis

3.1 Heatmap of Attributes Correlation with Rank

To better understand the relationship between song features and Billboard rankings, we created a heatmap showcasing the correlation coefficients for each feature. This visualization allows us to identify patterns and relationships at a glance, with the strength and direction of the correlations represented by varying shades of color. Darker shades indicate stronger correlations, whether positive or negative, making it easier to pinpoint which features are most closely associated with higher Billboard rankings.

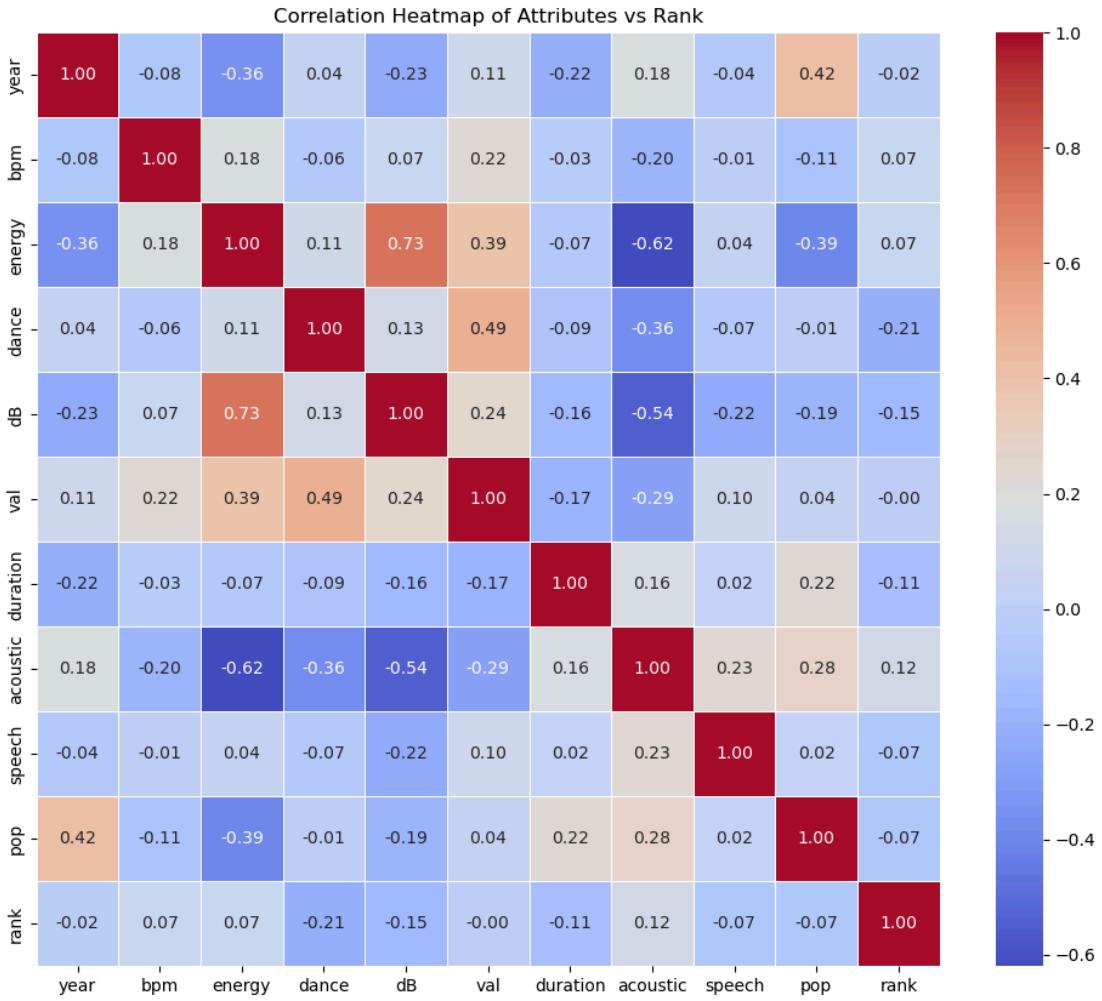


Figure 1 Correlation Heatmap of Attributes

The attribute most strongly correlated with *rank* (influence on song success) is **danceability** (correlation = 0.211), indicating that more danceable songs tend to rank higher. This is followed by **dB** (loudness) and **acousticness**, which also exhibit notable correlations with rank. Interestingly, **popularity** and **energy**, often assumed to impact song success, show weaker correlations with rank. Attributes like **speechness**, **bpm** (tempo), and **year** exhibit very low correlations, suggesting limited influence on rankings. This analysis emphasizes that certain audio features that allow listeners to dance and are loud may predict song success better than others.

3.2 Multiple Artists and Song Rank

We conducted an analysis to determine whether collaborations have an impact on a song's ranking on the Billboard charts and a song's popularity according to Kaggle. To do this, we categorized tracks into two distinct groups: solo performances, where a single artist is credited, and collaborations, which involve two or more artists working together. After grouping the songs accordingly, we calculated the average Billboard ranks for each category as well as the song's popularity to identify any significant differences. To visualize these findings, we created two box plot that compares the average rankings of solo tracks and collaborative tracks and average popularity, providing a clear representation of the potential influence of artist partnerships on rank and popularity performance.

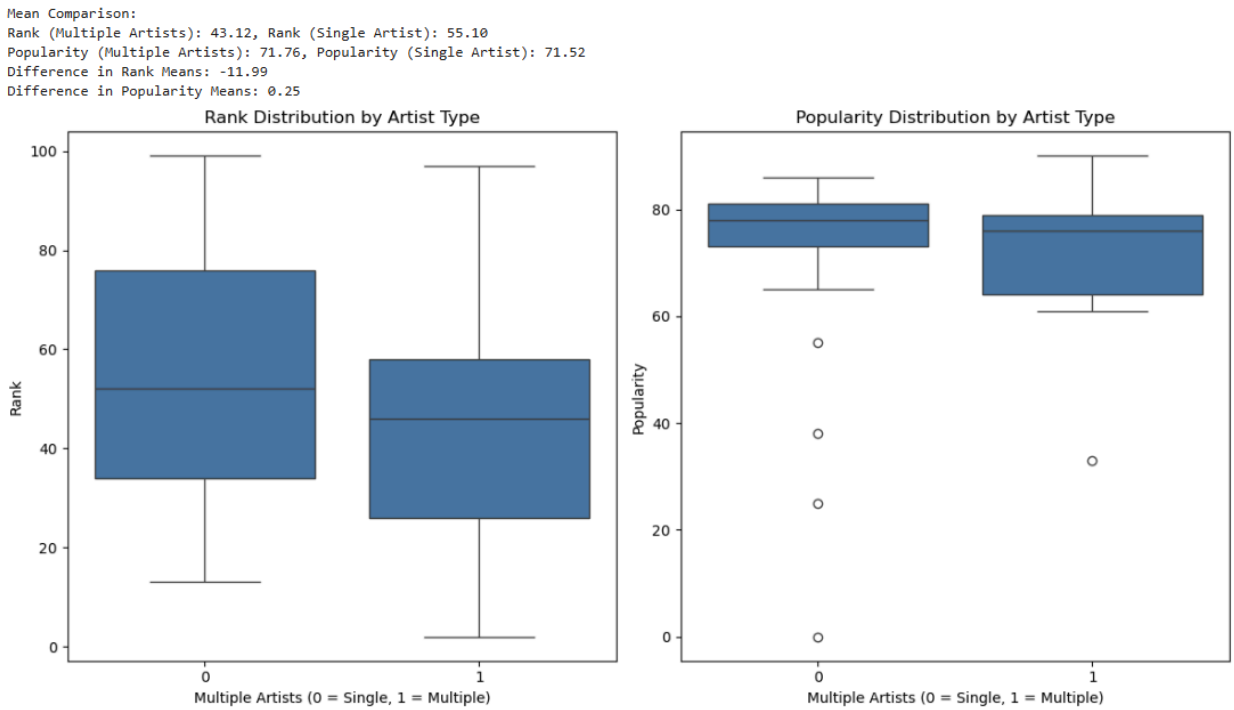


Figure 2 Rank and Popularity Distribution by Artist Type

Our analysis revealed that collaborations (**Multiple Artists**) significantly enhance a song's chances of achieving a better rank on Billboard's Hot 100. By collaborating multiple artists, there is a **12** step positive decrease in rank from 55 to 43 (lower ranking is better). However, for the popularity of a song, collaborating with multiple artists does not have a significant impact, having an increase of **0.25** in the opposite direction. By pooling fan bases and creative input, collaborations often generate broader appeal and higher commercial performance. This trend highlights the importance of strategic collaborations in the music industry because we are prioritizing the ranking of a song, merging the talents of multiple artists can achieve exceptional success with rankings.

3.3 Trends of Popularity and Rank Over the Decade

To thoroughly investigate the relationship between Spotify popularity scores and Billboard rankings, we utilized a scatterplot as a visual analysis tool and incorporated a trend line to enhance interpretability. We created these graphs to analyze the average Spotify popularity scores over the years and the average Billboard rankings over the same time period. By examining these trends side by side, we were able to compare how the average popularity of songs evolved over the years in relation to their average rankings on the Billboard charts. This comparison provides valuable insights into the relationship between streaming popularity and chart performance over time.

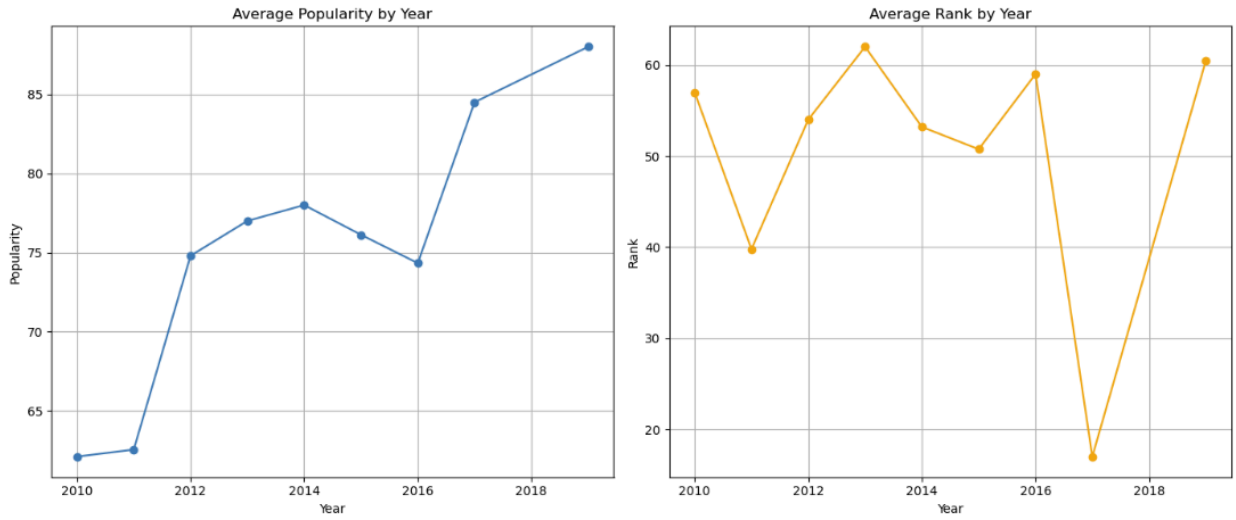


Figure 3 Rank and Popularity Over the Decade

The **Average Popularity by Year** graph (left) shows a steady increase in Spotify popularity scores, other than the slight dip in 2016, with scores peaking at the end of the decade, reflecting the growing influence of digital streaming on music consumption. In contrast, the **Average Rank by Year** graph (right) shows more variability in Billboard rankings, with fluctuations year to year and a notable dip in 2017, indicating stronger chart performance for that year's songs (as lower ranking numbers indicate higher ratings). While Spotify popularity has shown a clear upward trend over time, the relationship with Billboard rankings appears less consistent, suggesting that factors beyond streaming popularity, such as radio play, marketing, or physical sales, may also play a role in determining a song's chart success.

3.4 Energy Level and Popularity Across Genres

The next question we wanted to address was whether different energy levels in music correlated with specific genres that made a song popular. To explore this, we created a new column called energy_bin, which categorizes songs based on their energy levels. This column divides the energy scores into three bins: Low (0–49), Medium (50–74), and High (75 and above). With this categorization, we generated a bar chart to visualize which music genres were most popular within each energy level.

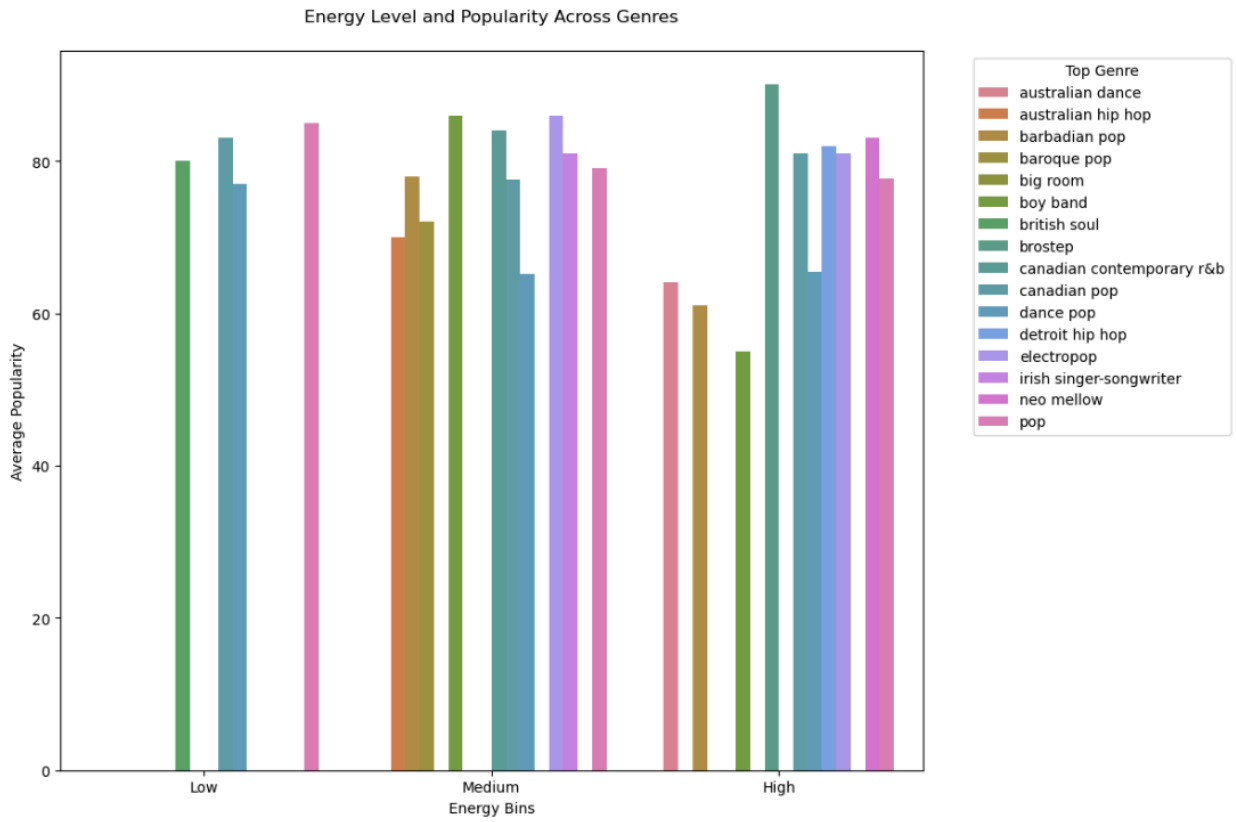


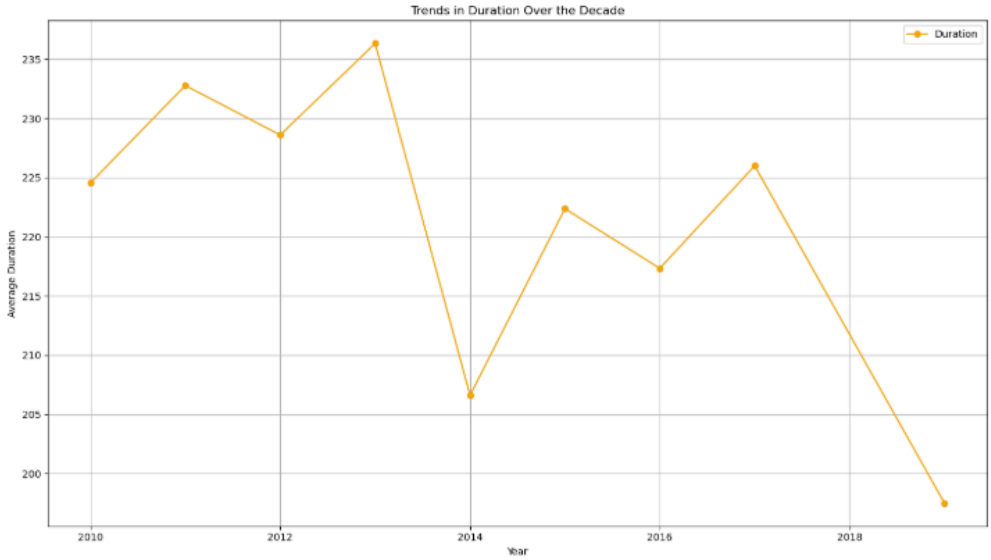
Figure 4 Genre by Energy Level

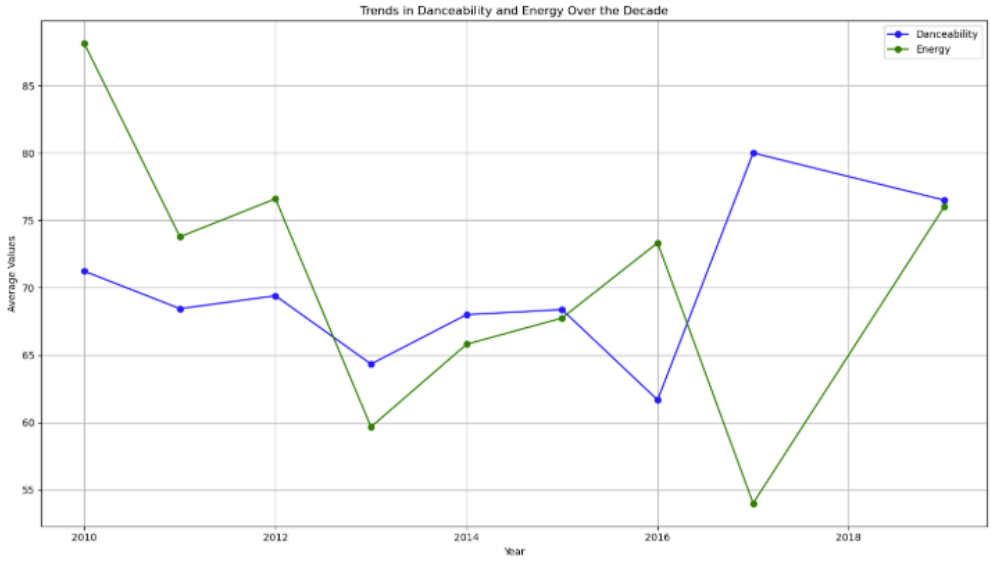
Pop and **Neo Mellow** emerge as the most frequently represented genres, maintaining high popularity across all three energy bins. Among these, **Pop** stands out as the most popular genre overall, consistently achieving high average popularity regardless of energy level. For clarification, an example of a Pop song would be *Shake It Off* by Taylor Swift, while Neo Mellow is represented by *Hey, Soul Sister* by Train. Some genres, such as **Big Room**, exemplified by Adele's *Hello*, show a significant rise in popularity at the "High" energy level, suggesting they benefit from higher energy. In contrast, niche genres like **Australian Hip Hop** (*Fancy* by Iggy Azalea Featuring Charli XCX) and **Barbadian Pop** (*Work* by Rihanna Featuring Drake) have lower popularity across most bins, indicating a smaller audience or less mainstream appeal. The presence of **Pop** and **Neo Mellow** across all three bins highlights their broad appeal, making them consistently popular choices. It is worth mentioning though that all three energy levels are shown in the top 100 songs of the decade, proving a diversified taste in the music industry.

3.5 Danceability, Duration, and Energy Trends Across the Decade

We set out to explore how Danceability, Duration (in seconds), and Energy have evolved in music between 2010 and 2019. Were there significant peaks or drops in these features, or did they remain relatively constant over time? Did any specific trends or popular styles emerge in certain years? To answer these questions, we created a time series chart to visualize the overall patterns in these three

characteristics. Additionally, we included an average chart for each feature by year to provide further insight into their yearly changes.





Average values by year:

	year	dance	duration	energy
0	2010	71.222222	224.555556	88.111111
1	2011	68.444444	232.777778	73.777778
2	2012	69.400000	228.600000	76.600000
3	2013	64.333333	236.333333	59.666667
4	2014	68.000000	206.600000	65.800000
5	2015	68.375000	222.375000	67.750000
6	2016	61.666667	217.333333	73.333333
7	2017	80.000000	226.000000	54.000000
8	2019	76.500000	197.500000	76.000000

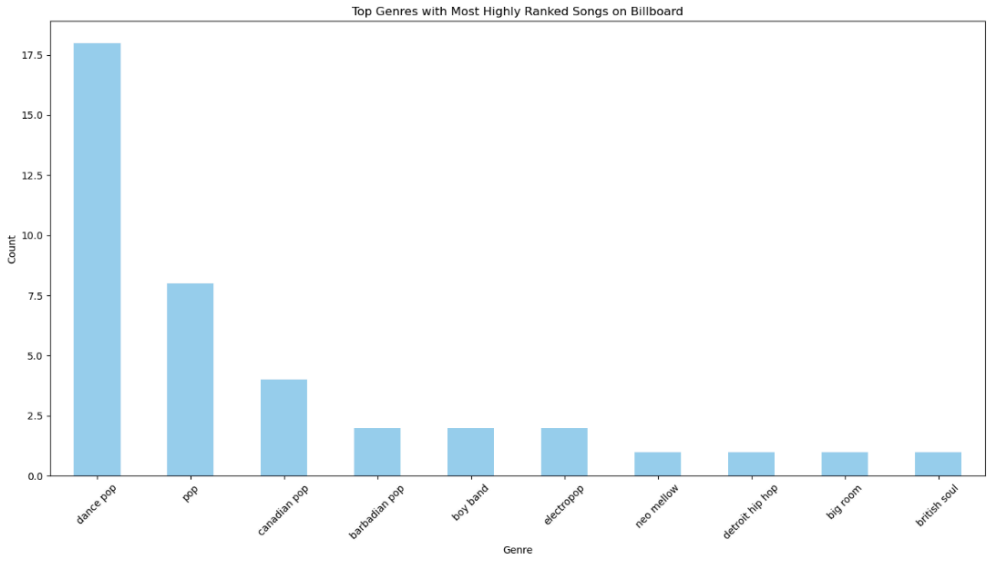
Figure 5 Time Series Data on Danceability, Duration, and Energy

The graphs highlight trends in **Duration**, **Danceability**, and **Energy** of songs from 2010 to 2019. The top chart shows that the average song duration has consistently declined over the decade, peaking in 2013

at around 235 seconds before experiencing a sharp drop in 2014 to around 205 seconds. By 2019, duration ended at approximately **197.5 seconds**, marking a significant decline from earlier years. The bottom chart reveals that **Danceability** remains relatively stable, fluctuating between 65 and 75, with a slight increase toward the end of the decade, ending in 2019 at approximately **76.5**. **Energy**, on the other hand, shows dramatic fluctuations, starting high in 2010 (~88), dropping significantly by 2014 (~59), and recovering inconsistently to end at **76** in 2019. The shortening of songs likely reflects a shift toward more concise, streaming-friendly tracks that cater to modern listening habits, where shorter content tends to perform better. The stability in **Danceability** suggests that rhythm and movement remain key features of popular music, while the fluctuations in **Energy** may indicate evolving production styles and audience preferences, with trends in genres or moods influencing the intensity of music over the decade. Overall, these changes illustrate how the music industry adapts to technological advances and shifting consumer demands.

3.6 Most Ranked Genre Throughout the Decade

What genre or genres had the most highly ranked songs on Billboard? To find out, we created a simple graph to highlight the most popular genres of the 2010s. This information gives us insight into what music resonated with consumers during that decade and could even reflect key cultural moments in history.




```
Top genres by count:
top_genre
dance pop                    18
pop                           8
canadian pop                  4
barbadian pop                 2
boy band                      2
electropop                    2
neo mellow                    1
detroit hip hop               1
big room                      1
british soul                  1
baroque pop                   1
australian hip hop            1
canadian contemporary r&b     1
irish singer-songwriter       1
australian dance              1
brostep                       1
Name: count, dtype: int64
```

Figure 6 Most Common Genres Between 2010-2019

The chart highlights the genres with the most highly ranked songs on Billboard during the 2010s. **Dance Pop** dominates with 18 songs, making it the clear leader, followed by **Pop** with 8 songs and **Canadian Pop** with 4 songs. These results suggest that upbeat and mainstream pop music resonated most with audiences during this decade. Smaller genres like **Barbadian Pop**, **Boy Band**, and **Electropop** appear with 2 songs each, while niche genres such as **Neo Mellow**, **Detroit Hip Hop**, and **Big Room** are represented by only 1 song. This distribution reflects the 2010s' cultural preference for high-energy, danceable tracks, with global artists like Canadian and Barbadian performers contributing significantly. The dominance of Dance Pop suggests its widespread appeal and influence during this period, marking a pivotal moment in shaping the decade’s music trends.

4. Conclusion

In this project, we analyzed factors influencing song success, focusing on audio attributes, collaborations, genre trends, and the evolution of music over the decade. From the analysis questions presented in the proposal, we found the following results:

1. **Is there a strong correlation between song features and Billboard rankings?**

   Danceability showed the highest correlation (0.211) with rankings, followed by loudness and acousticness, while tempo, energy, and year had negligible impact.

2. **Do collaborations influence Billboard rankings?**

   Collaborative tracks consistently ranked higher, emphasizing the importance of combining fan bases and creative strengths for broader appeal.

3. **What is the relationship between Spotify popularity and Billboard rankings?**

   While Spotify popularity grew over the decade, its relationship with rankings was inconsistent, with factors like radio play and marketing likely playing significant roles.

4. **Do energy levels correlate with specific genres?**

   Big Room and Dance Pop thrived in high energy engineering only, while Neo Mellow showed appeal across all energy levels, highlighting genre flexibility.

5. **How have song characteristics evolved over the decade?**

   Song durations decreased significantly, reflecting streaming-friendly trends. Danceability remained stable, while energy levels fluctuated, showing shifts in production styles and audience preferences.

6. **Which genres dominated Billboard rankings in the 2010s?**

   Dance Pop dominated with 18 highly ranked songs, followed by Pop and Canadian Pop, reflecting the mainstream appeal of high-energy genres.

This project faced a few limitations, including a small dataset and the lack of non-music factors like marketing. This dataset also became a part of a Subscription only usage, and we were unable to properly scrape directly from the site because of this. Future work could integrate broader datasets, explore lyrical content, and examine regional trends for a more comprehensive understanding of song success.