Exploring Music Trends and Popularity Factors

By Samuel Warren and Anshul Sadh-Gauri

May, 2023

Project Overview

Songs on music platforms often skyrocket to popularity seemingly out of nowhere. Take, for instance, the phenomenon of "Old Town Road." While established artists may effortlessly amass millions of plays, emerging artists often aspire for that one breakthrough song. What factors contribute to the sudden popularity of certain songs? In this project, we aim to scrutinize data from various songs throughout history to uncover commonalities among popular tracks and examine whether these characteristics have evolved over time. The data will be analyzed to help shed light on the below questions:

- How do factors such as Genre, BPM, Length, Decibel level, and Speechiness influence a song's popularity?
- Are specific levels of these characteristics more prevalent during certain historical periods?

Hypothesis: Commencing from 1956, and progressing towards 2019, we anticipate observing a reduction in song length, an increase in BPM, and a consistent decibel level.

This project only utilizes basic music concepts and does not delve into intricate musical theories. Consequently, all readers should find the project accessible and comprehensible. The only potential source of confusion may arise from the variables in the datasets, which will be clarified in the subsequent section.

Data Sources

Although we explored a number of different open datasets, including the Top 100 Songs on Spotify, the project primarily used SP2000, which encompasses data and common characteristics of approximately 2,000 popular songs spanning the years 1956-2019.

SP2000 was sourced from a Kaggle user who compiled 2,000 songs and processed them through the Spotify API. This API provides detailed characteristics for each song. It's important to note that this compilation was last updated in 2019, hence including songs only up to that year.

It should also be noted that the Kaggle contributor focused on evaluating the Spotify API and creating music recommendation software, whereas the current project is centered on exploring trends within the dataset.

Data Analysis

Part 1 - Data Cleansing

In its original state, SP2000 comprised 15 columns. To streamline our analysis, we opted to remove eight columns and introduce one new column. The Top Genre column was modified to include only basic genres, amalgamating similar genres like alternative rock and classic rock under 'rock.' Additionally, genres that didn't align with a specific category were consolidated into the 'Other' column, except 'Dutch Cabaret,' which remained a distinct category due to its 51 associated songs.

The newly added column assigns a numeric value ranging from 1 to 63, corresponding to the years 1958 to 2019. This numeric representation facilitates the use of a Linear Model, allowing for a more generalized selection of the year value, rather than requiring an exact year.

Next, we categorized each variable as either numerical or categorical:

- **Index**: This variable is a unique value for each row in the data set.
  - Data Type: factor
  - Range/Levels: 1-1994

- **Title**: This variable is the title of the song.
  - Data Type: categorical
  - Range/Levels: unique title

- **Artist**: This variable is the artist of the song.
  - Data Type: categorical
  - Range/Levels: unique artist

- **Top Genre**: This variable indicates the music genre of the song. The genre has been generalized.
  - Data Type: numeric
  - Range/Levels: adult standards, country, Dutch cabaret, electric/dance, folk, funk, g funk, hip hop, indie, jazz/blues, metal, other

- **Year**: This variable records the year when the song was released.
  - Data Type: categorical
  - Range/Levels: 1956-2019

- **Beats Per Minute (BPM)**: This variable measures the pace of a song/how fast a song is. Specifically, it records the amount of beats a song has in one minute.
  - Data Type: numeric
  - Range/Levels: 37-206

- **Loudness (dB)**: This variable measures the loudness of a song. This is represented as a ratio in comparison to other songs on Spotify. In this instance, a dB level of -2 represents

the loudest in our data set, and -27 represents the lowest dB level.
  - **–** Data Type: numeric
  - **–** Range/Levels: -27-(-2)

- **Speechiness**: This variable records the number of words in a song. This number is then compared as a ratio to the other songs on the playlist.
  - Data Type: numeric
  - Range/Levels: 2-55

- **Popularity**: This variable records the popularity of a song. This is calculated by combining a few different factors such as the number of plays, user engagement (shares/likes), and the percentage of the song listened to on average. The final value is a ratio from 1-100.
  - Data Type: numeric
  - Range/Levels: 1-100

- **Length (Duration)**: This variable represents the length of a song in seconds.
  - Data Type: numeric
  - Range/Levels: 93-966

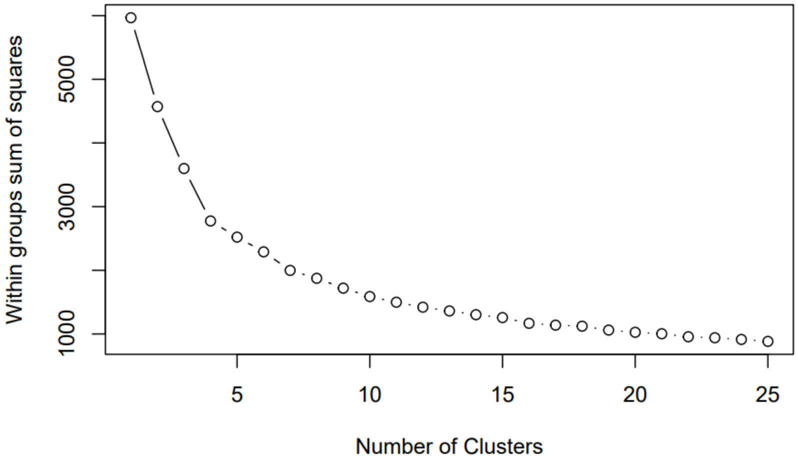
- **YearNumber**: This variable is a number corresponding to the year. It is used in models and plots.
  - Data Type: numeric
  - Range/Levels: 1-63

Part 2 - Cluster Analysis (k-means) and Data Visualization

The first model we ran was a cluster sampling with k-means. The objective is to ascertain if variables are grouped by year, indicating a potential correlation. To begin, we created a data frame with specific variables, namely loudness (dB), song length, and beats per minute (BPM). Post-creation, we scaled the data to facilitate k-means analysis.

Subsequently, a function was employed to generate a WSS (Within Sum of Squares) plot, aiding in determining the optimal number of clusters. The WSS plot indicated that the optimal number of clusters was six, and with this information, we used k-means to cluster the data.

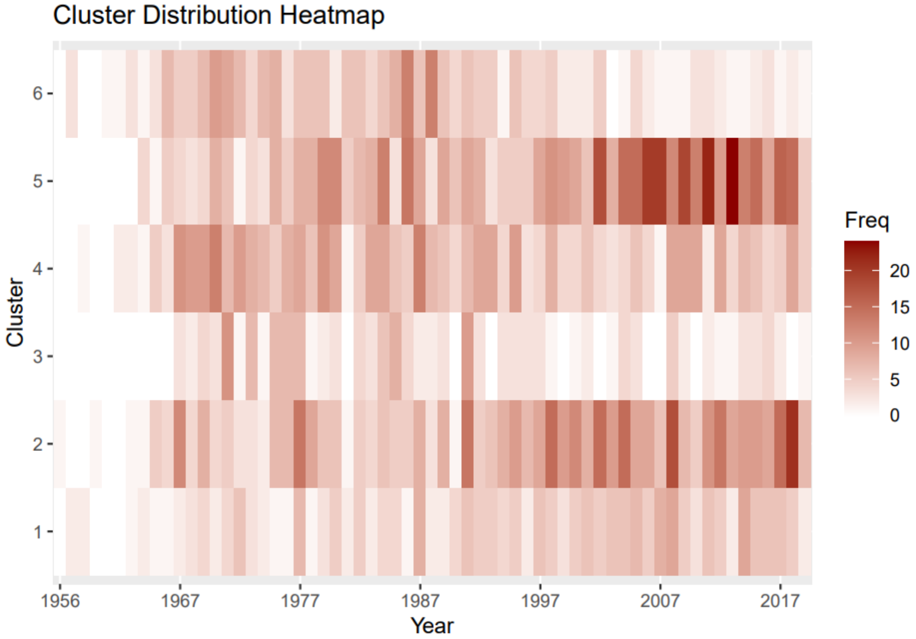
Following the k-means analysis, each cluster encompassed a balanced representation of a few hundred songs.



The next step involved creating visual plots to elucidate cluster patterns and identify potential trends.

Cluster Distribution Heatmap

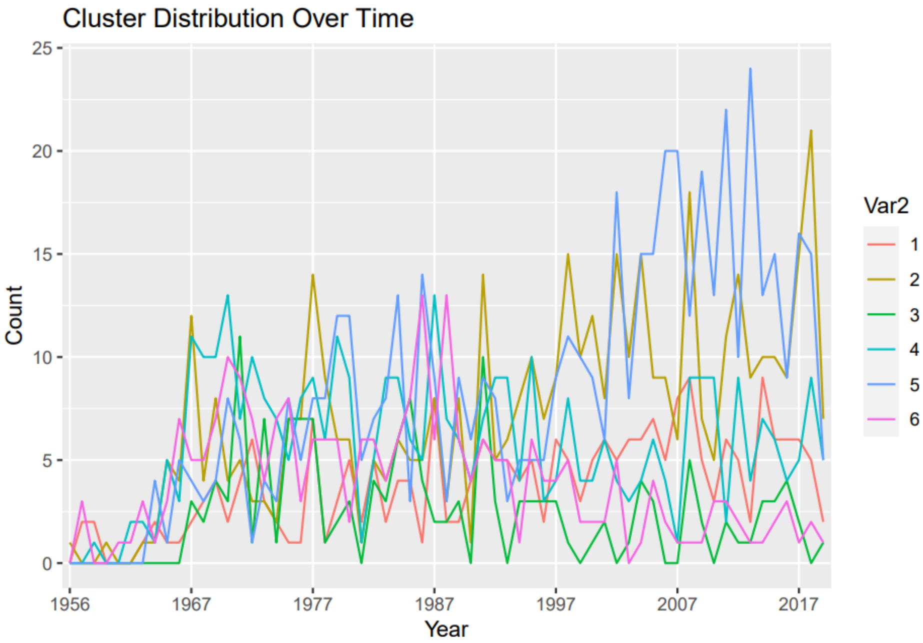
The heatmap effectively illustrates the trends within each cluster, showcasing the similarity among songs within each group. Notably, the songs constituting each cluster share common values for BPM, Decibel Level, and Length. Over time, Clusters 1, 2, and 5 show an increase in frequency, while Clusters 6 and 3 decrease. Cluster 4 maintains a relatively constant frequency.



Cluster Distribution Over Time

The line plot shown below provides an insightful visualization of cluster trends. In alignment with the heatmap, Clusters 1, 2, and 5 show an increase over time, while Clusters 3 and 6 decrease, and Cluster 4 remains constant.

Overall, based on these two plots, it becomes evident that song variables, or song characteristics, change over time. It is unlikely for a cluster, especially five of them, to center around a specific year without a connection to time. Given the concentration of clusters over certain periods, there is a clear association with temporal progression.

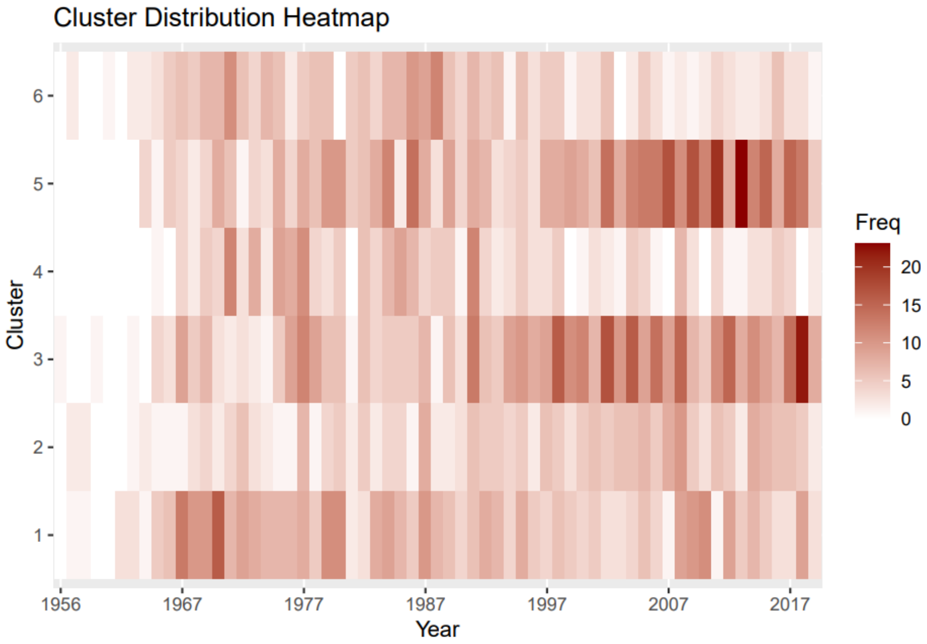


Part 3 - Cluster Analysis (PAM Clustering) and Data Visualization

While k-means excels in determining data clusters, it is crucial to employ multiple models to account for errors. The subsequent step involves implementing a PAM clustering method.

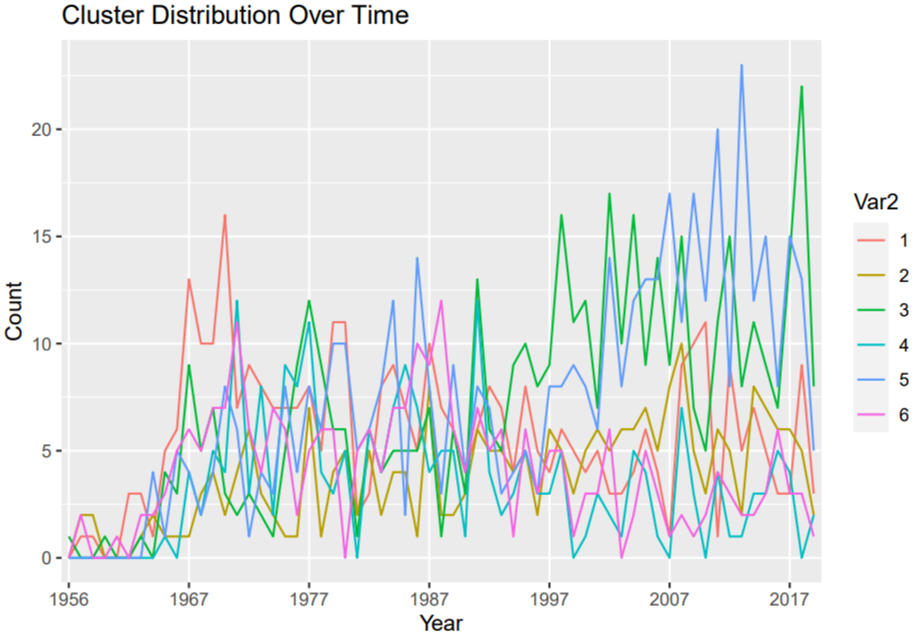
Cluster Distribution Heatmap

The heatmap generated by the PAM clustering method provides a comparable representation to k-means. Similar to k-means, there is an increase in frequency as time approaches 2019 in three clusters. Among the remaining three clusters, two exhibit a decrease in frequency over time, while the last remains relatively constant.



Cluster Distribution Over Time

Once again, the line plot aligns with the patterns observed in the previous plots.



In summary, both the PAM and k-means clustering models underscore the relationship between time and specific values of variables associated with Length, Decibel Level, and BPM. With all but one cluster concentrated over specific time frames, songs from these periods are distinguishable based on their BPM, length, and decibel level.

Part 4 - Linear Regression and Data Visualization

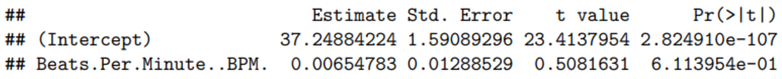
With the established understanding that specific levels of BPM, decibel level, and length correspond to certain times, the next step involved running simple linear regression models.

Univariate Models

The model was designed to predict YearNumber, a numerical variable ranging from 1 to 63, with BPM, decibel level, and length as individual predictors.

Model Predictor: Beats Per Minute

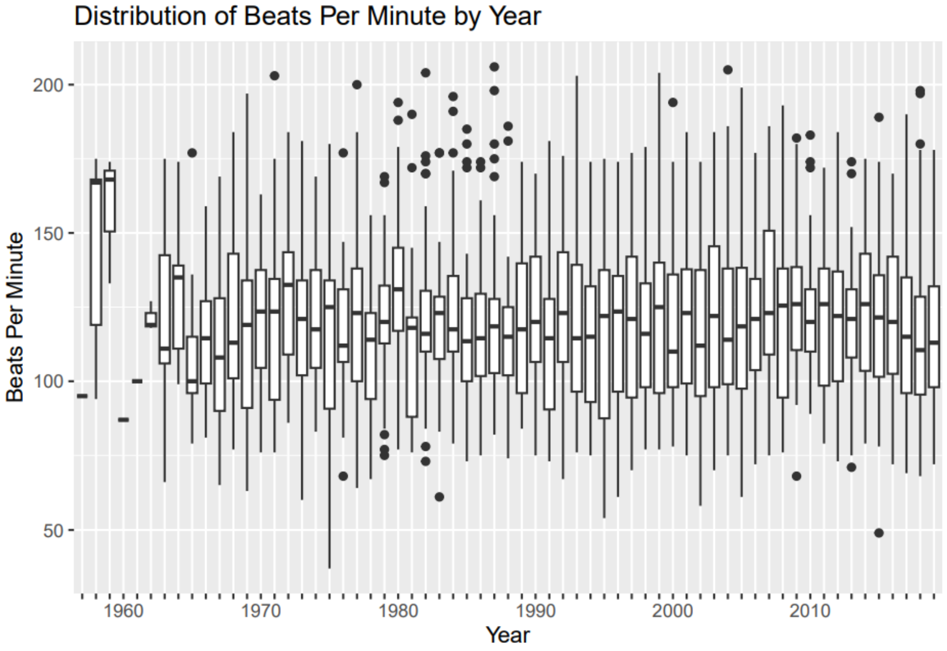
A linear model was executed with beats per minute as the predictor.



```
##                             Estimate Std. Error    t value      Pr(>|t|)
## (Intercept)              37.24884224 1.59089296 23.4137954 2.824910e-107
## Beats.Per.Minute..BPM.    0.00654783 0.01288529  0.5081631  6.113954e-01
```

Since the p-value is 0.611 is not significant, we concluded that the beats per minute of a song stay constant over time.

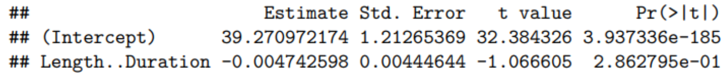
Here is a plot that shows the average beats per minute over time.



This plot confirms the above statement. Over time, the beats per minute of a song, on average, stay roughly constant.

Model Predictor: Length

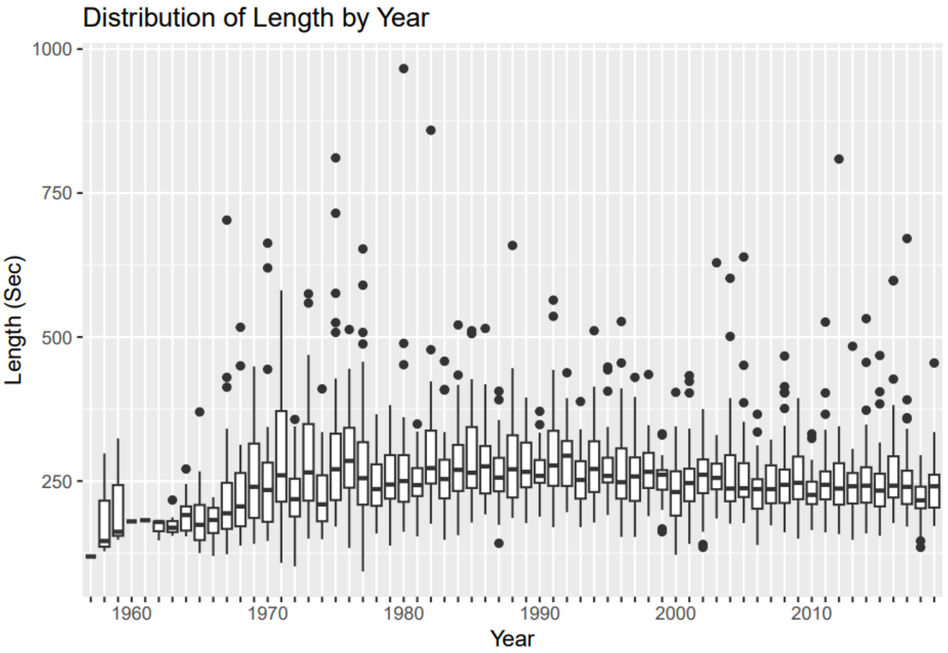
Next, we ran a model with Length as the predictor.



```
##                      Estimate Std. Error   t value      Pr(>|t|)
## (Intercept)      39.270972174 1.21265369 32.384326 3.937336e-185
## Length..Duration -0.004742598 0.00444644 -1.066605  2.862795e-01
```

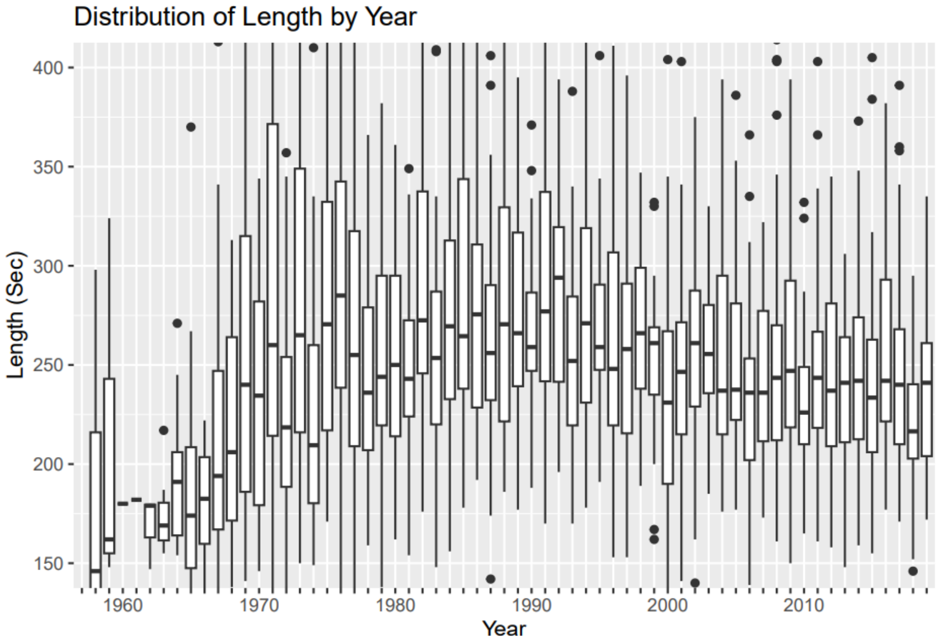
The p-value is greater than 0.05 and as such is not significant. Thus we conclude that Song Length stays relatively constant over time.

Here is a plot that shows the average length over time.



Upon closer examination, this plot challenges the initial conclusion above. While the average length doesn't exhibit drastic increases or decreases after 1970, it notably peaked during the 1980s and 1990s. Subsequent years in the 2000s and 2010s show a slightly lower average length.

The following plot provides a more focused view of the trends:

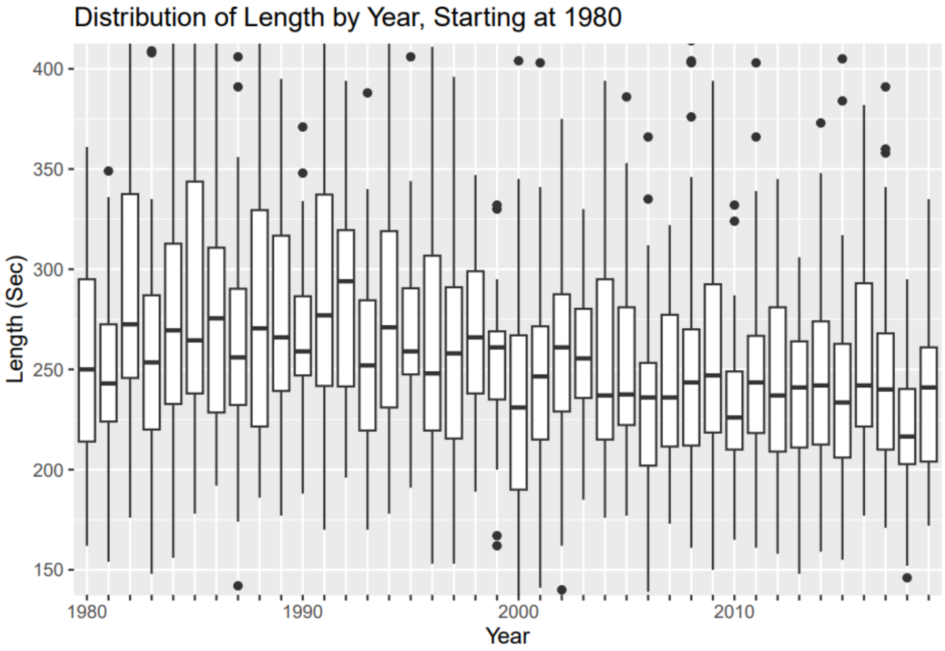


The plots show that the length of a song changes over time. Songs had a higher average length in the 1980s and 1990s, followed by a decrease during the 2000s and 2010s. We attribute this difference to the non-linear distribution of length by year, which experiences fluctuations over the years 1956-2019. A linear model, beginning in 1980 and ending in 2019, shows an overall decrease in length.

```
##                     Estimate  Std. Error   t value      Pr(>|t|)
## (Intercept)       53.0920048 1.093605101 48.547693 7.287135e-308
## Length..Duration  -0.0291609 0.004007705 -7.276208  5.562155e-13
```

Based on the summary, the p-value is less than 0.05. Given this and the negative coefficient, it is evident that song length has decreased since 1980.

Below is an additional plot to visualize this trend:

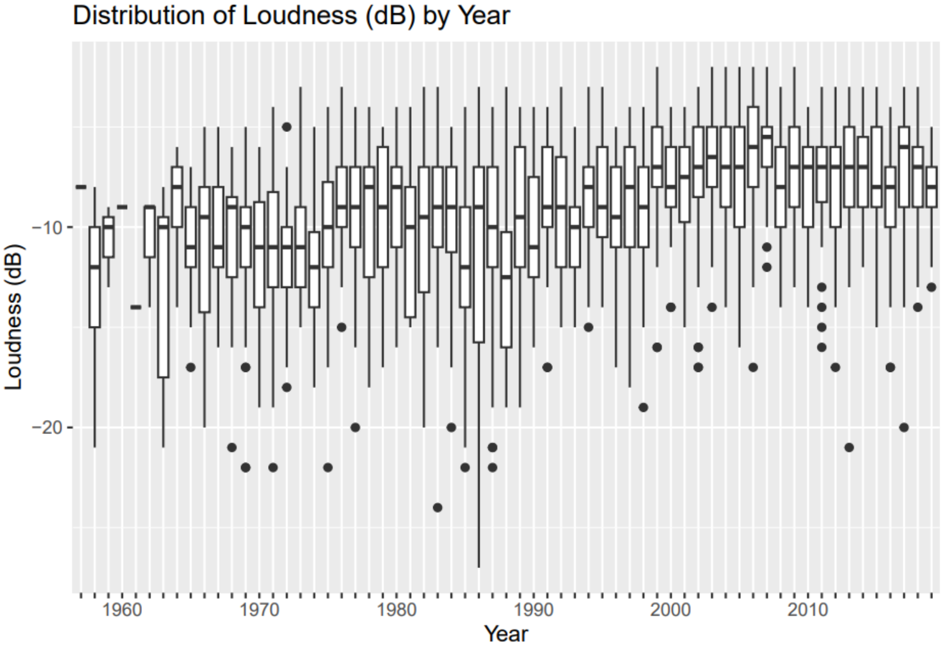


Model Predictor: Decibel Level

Next, we explored the decibel level over time.

```
##                 Estimate Std. Error  t value     Pr(>|t|)
## (Intercept)    51.677111 0.90437058 57.14152 0.000000e+00
## Loudness..dB.   1.515405 0.09313468 16.27111 5.563034e-56
```

The decibel level appears to be significant ($p < 0.05$). Consequently, since the coefficient is positive, the decibel level of a song increases over time. The following plot confirms our conclusions:



Except for a few outlier years, particularly around the late 1980s, the decibel level of a song consistently increases over time, approaching 2019. Disregarding outliers, the lowest point occurred in the early 1970s, while the peak was observed from 2006 to 2019.

These plots contribute to addressing our guiding question: Are certain levels of these characteristics more common at specific times in the past?

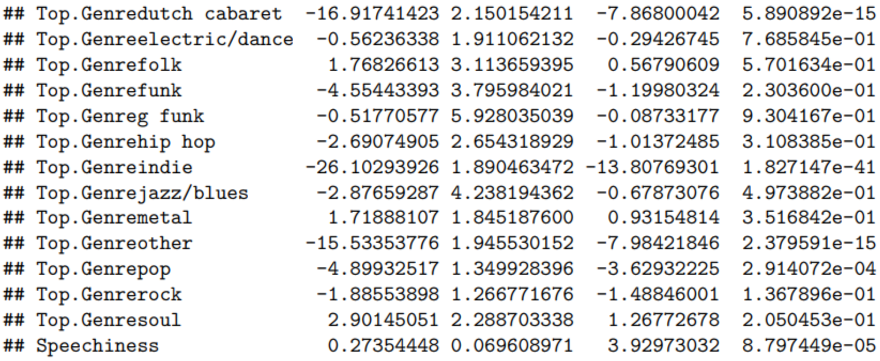
In summary, insights from the plots indicate that certain variables change over time. Beats Per Minute remains relatively constant, with the average value per year hovering around 120. Song length experiences fluctuations, starting at approximately 180 seconds or 3 minutes, peaking in the 80s and 90s at around 270 seconds (3 minutes and 30 seconds), followed by a decrease in average length to around 250 seconds today, equivalent to 4 minutes and 10 seconds.

Moreover, the decibel level consistently increases over time. Its average in 1956 was around -10 dB on the Spotify API scale, steadily rising to a peak at around -7 in 2006. This level has remained constant from 2006 to 2019.

<u>Multivariate Models</u>

The next step was to run a model predicting a song's popularity based on the multiple predictors: Length, BPM, Decibel level, Genre, and Speechiness.

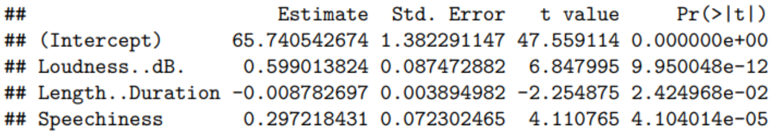
```
##                             Estimate  Std. Error     t value      Pr(>|t|)
## (Intercept)              73.05778183 2.131528712 34.27482886 2.358493e-202
## Beats.Per.Minute..BPM.   -0.01404656 0.010421775 -1.34780867  1.778748e-01
## Loudness..dB.             0.54058674 0.084093776  6.42837989  1.613340e-10
## Length..Duration         -0.01688194 0.003668022 -4.60246511  4.441195e-06
## Top.Genrecountry         -9.49558846 3.897984542 -2.43602517  1.493770e-02
```



```
## Top.Genredutch cabaret  -16.91741423 2.150154211  -7.86800042  5.890892e-15
## Top.Genreelectric/dance  -0.56236338 1.911062132  -0.29426745  7.685845e-01
## Top.Genrefolk             1.76826613 3.113659395   0.56790609  5.701634e-01
## Top.Genrefunk            -4.55443393 3.795984021  -1.19980324  2.303600e-01
## Top.Genreg funk          -0.51770577 5.928035039  -0.08733177  9.304167e-01
## Top.Genrehip hop         -2.69074905 2.654318929  -1.01372485  3.108385e-01
## Top.Genreindie          -26.10293926 1.890463472 -13.80769301  1.827147e-41
## Top.Genrejazz/blues      -2.87659287 4.238194362  -0.67873076  4.973882e-01
## Top.Genremetal            1.71888107 1.845187600   0.93154814  3.516842e-01
## Top.Genreother          -15.53353776 1.945530152  -7.98421846  2.379591e-15
## Top.Genrepop             -4.89932517 1.349928396  -3.62932225  2.914072e-04
## Top.Genrerock            -1.88553898 1.266771676  -1.48846001  1.367896e-01
## Top.Genresoul             2.90145051 2.288703338   1.26772678  2.050453e-01
## Speechiness               0.27354448 0.069608971   3.92973032  8.797449e-05
```

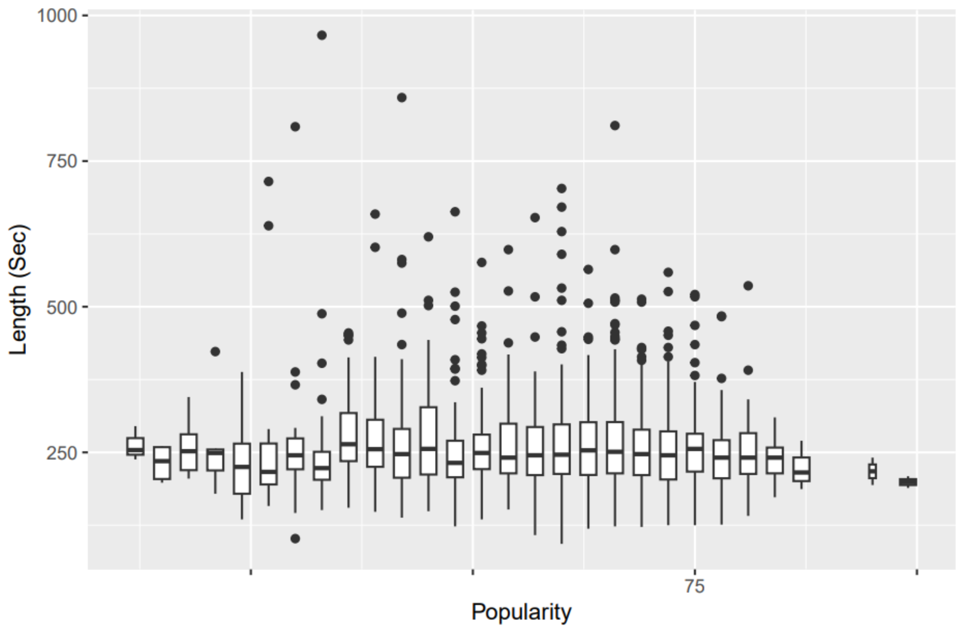
Upon executing the model, not all variables showed significance and, consequently, did not contribute significantly to popularity. Specifically, BPM was found to be statistically insignificant. An examination of Genres revealed that only specific genres influenced overall popularity. The insignificance of Genre becomes apparent as the majority of genres register above the p-value, with those below exhibiting a low sample size. The exception to this trend is the "pop" category, which stands out as statistically significant. Pop is the most popular genre, a conclusion supported by the sheer volume of data. Pop comprises 435 of the 1990 songs, making its popularity more pronounced than other categories.

The subsequent step involved running the model using only these three variables. Additionally, we generated three graphs comparing each variable to popularity individually.

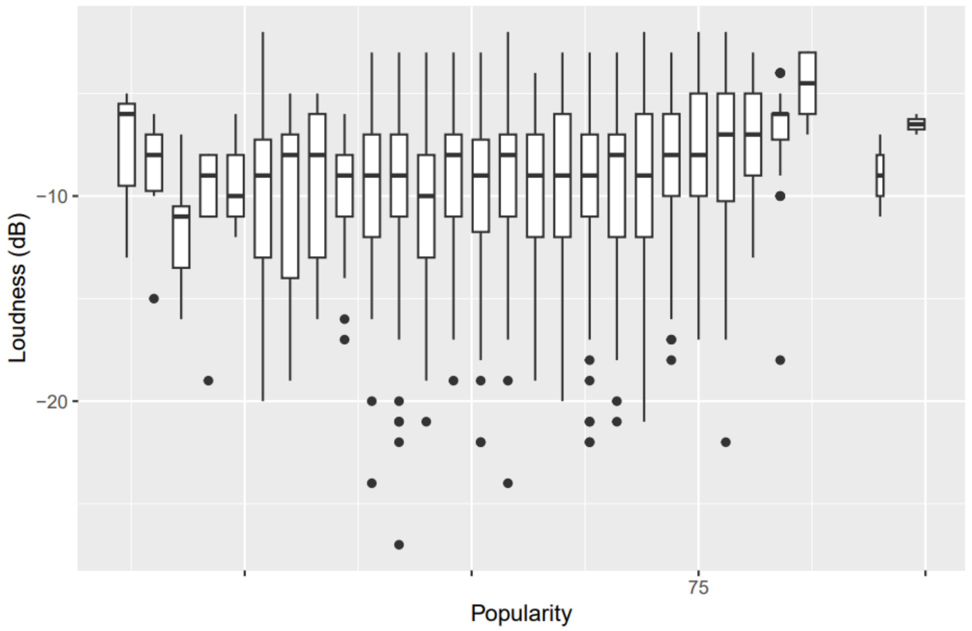


```
##                     Estimate  Std. Error   t value     Pr(>|t|)
## (Intercept)      65.740542674 1.382291147 47.559114 0.000000e+00
## Loudness..dB.     0.599013824 0.087472882  6.847995 9.950048e-12
## Length..Duration -0.008782697 0.003894982 -2.254875 2.424968e-02
## Speechiness       0.297218431 0.072302465  4.110765 4.104014e-05
```

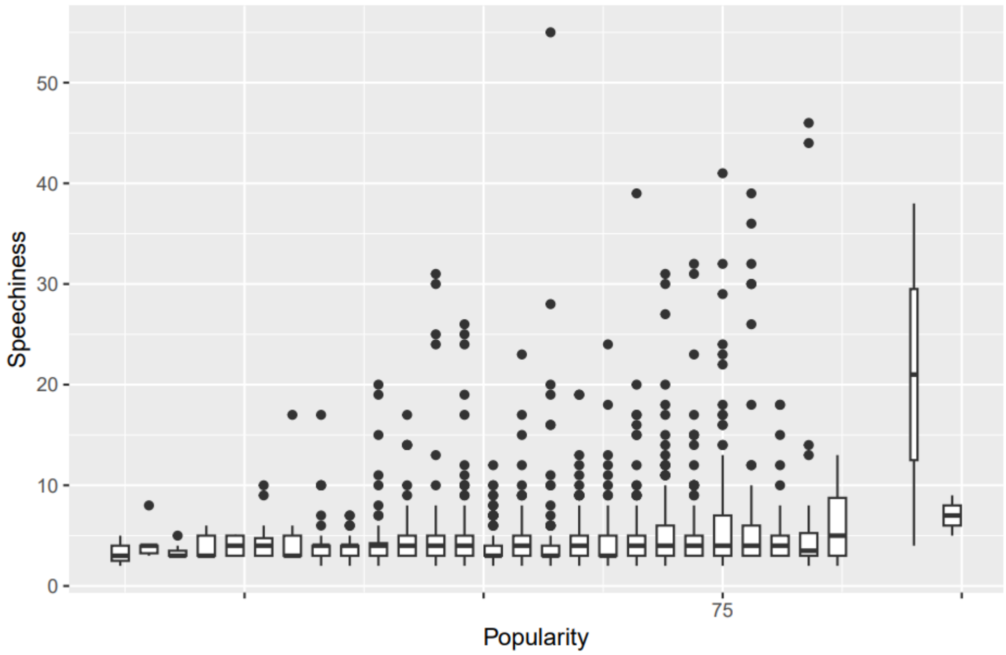
Distribution of Length by Popularity



Distribution of Loudness by Popularity



Distribution of Speechiness by Popularity



Based on the new model, all of the variables were found to be significant (p-value < 0.05). Analysis of the model and examination of the two plots revealed that the Length variable has minimal to no effect on popularity. On average, an eight-minute-long song is less popular than a two-minute one. However, confirming this is challenging due to the preponderance of songs below 4 minutes (240 seconds), and the scarcity of longer songs. The p-value for length is less than 0.05, signifying significance. Yet, the associated coefficient is merely -0.0087. This implies that for every extra second, popularity decreases by less than 0.01 on a scale of 0-100. Based on this dataset, the conclusion is that song length does not significantly impact popularity.

Conversely, the decibel level appears to exert a slight influence on popularity. Despite a concentration of songs around -5 decibels, the positive slope suggests significance. The p-value is highly significant at <.0001, and the coefficient is substantial at 0.5990. As all decibel values are negative, the inference is that, on average, louder songs tend to be more popular. The average of songs with a popularity rating above 75 is slightly greater than those with a lower rating. In conclusion, louder songs tend to be more popular according to this dataset.

Lastly, the Speechiness value seems to moderately affect popularity. On average, the Speechiness value hovers around 4-5 for most popularity levels. Notably, when popularity is very high, Speechiness values exhibit numerous outliers. Many data points around or above 75 popularity have Speechiness values significantly higher, reaching 20 or 30. Overall, it can be inferred that Speechiness has a modest impact on popularity, where an increase in Speechiness corresponds to an increase in song popularity. It is important to note that the majority of songs have low Speechiness values, contributing somewhat significantly to popularity.

Summary and Conclusions

The data analysis provided valuable insights into our guiding questions: a) How do factors such as Genre, BPM, Length, Decibel level, and Speechiness influence song popularity? and b) Are certain levels of these characteristics more common at specific times in the past?

Variables Affecting Popularity:

The analysis indicates that only specific variables significantly impact song popularity. Decibel Level, Song Genre, and Speechiness emerge as influential factors. While Decibel level and Speechiness contribute to a minor boost in popularity, altering a song's genre to pop or indie styles can significantly enhance its appeal. The ideal combination for a potentially popular song would be a wordy, average-paced track at 120 BPM, characterized by louder pop elements. These findings lay the foundation for the exploration of the "perfect song," with ample room for further investigation into popular lyrical topics and additional musical elements.

Temporal Trends:

Addressing our second guiding question, Beats Per Minute remains relatively constant over time, with an average value hovering around 120 BPM. However, the length of songs evolves, starting at approximately 180 seconds or 3 minutes and experiencing an increase until the 80s and 90s, peaking at around 270 seconds or 4 minutes 30 seconds. Subsequently, there is a decline in average song length to approximately 250 seconds today, equivalent to 4 minutes and 10 seconds.

Decibel Level Changes Over Time:

Decibel level exhibits a constant increase over time. Starting with an average of around -10 dB in 1956 on the Spotify API scale, it steadily rises to peak at around -7 dB in 2006. This level remains consistent from 2006 to 2019. The Spotify API and the extensive history of songs offer abundant data for further exploration.

Future Directions

For future iterations of this project, expanding the dataset by including more music could solidify conclusions. Additionally, delving into the analysis of lyrics and musical notes would provide further depth and uncover commonalities. The potential for continuous exploration and discovery remains vast with the Spotify API and the rich history of songs as valuable resources.

References

1. Spotify Top 2000s Mega Dataset: This dataset was contributed by a Kaggle user who focused on evaluating the Spotify API and creating music recommendation software.
2. Top 100 Most Streamed Songs on Spotify: This dataset provides information about the most streamed songs on Spotify as of 2019.
3. CSV Data: Access to the CSV data used in our analysis is available on Google Drive.
4. R Documentation: This guide was used to develop the methodology and implement

analytic techniques.

5. Documentation for ggplot2: This guide on ggplot2 was used for creating visualizations that effectively communicated our results.
6. A Comprehensive Guide on ggplot2 in R: This guide from Analytics Vidhya provided additional insights and tips on using ggplot2 for data visualization.
7. k-Medoids in R: Algorithm and Practical Examples: This resource from Datanovia was referenced for understanding and implementing the k-Medoids clustering algorithm in R.
8. ChatGPT: Consultation on inquiries related to ggplot usage and interpretation of models and plots.