Name: Robert Mahoney

Date: 5/5/2025

Course: CIS366.NO1

Instructor: Jiaying Liu

Most Popular Spotify Songs

Introduction

In the age of streaming, understanding what makes a song popular on platforms like Spotify is valuable for artists, producers, and data analysts alike. This report explores the question, what makes songs popular? By identifying which musical characteristics, such as energy, danceability, or valence, are most strongly associated with high popularity scores, I aim to uncover patterns that explain listener preferences. To answer this, we will clean and prepare a Spotify dataset and then perform exploration data analysis to investigate how track features relate to popularity, ultimately revealing the traits that define a hit song.

Packages Used

To perform this analysis, I used several R packages that streamline data cleaning, manipulation, and visualization. The tidyverse package is a core collection that includes tools like dplyr for data wrangling, ggplot2 for data visualization, and readr for importing CSV files. I also used lubridate to handle and normalize date values, allowing me to extract the release year of songs regardless of inconsistent formatting. The stringr package was employed for cleaning character strings, such as extracting the year from text-based date formats. Together, these packages enabled efficient transformation of raw Spotify data into a tidy and analyzable format.

Data Preparation

For this project, I used the spotify_songs.csv dataset, which includes detailed information on songs featured in various Spotify playlists. The dataset originally contained 23 variables describing track metadata, audio features, and playlist information. Upon initial inspection, we identified 5 missing values in each of the track_name, track_artist, and track_album_name columns. Since these fields are essential identifiers for a song, I

chose to remove the corresponding rows to preserve the integrity of my analysis rather than impute incomplete song records.

To prepare the dataset for analysis, we cleaned several variables to ensure consistency. The track_album_release_date column had inconsistent date formats, some entries were complete (YYYY-MM-DD), while others listed only the year. Using the lubridate and stringr packages, I extracted the year only from this column and created a new release_year variable for uniformity and easier grouping in analysis. I also converted duration_ms to minutes and created a new column duration_min, which is easier to understand and read. Another choice I made was to remove songs with a duration of less than 30 seconds. This ensures actual songs are being considered in the analysis and not sounds or jingles. I also standardized all character fields by converting to lowercase and removing unnecessary punctuation or whitespace, ensuring consistent grouping and filtering. The column instrumentalness was originally stored as numeric values but occasionally appeared in scientific notation, like 0.00e+00, so I formatted these values to standard decimals and, rounded to 3 digits, for easier interpretation and visualization. Lastly, for every numeric data type I ensured they aligned with the data dictionary's ranges and filtered out the rest to ensure an accurate analysis.

A summary of the cleaned data set can be seen here which describes each variable, shows how each variable is of the right type, and is in line with the data dictionary:

Variable	**Type**	**Summary**
track_id	Character	Unique identifier for each track
track_name	Character	Name of the track
track_artist	Character	Artist performing the track
track_popularity	Numeric	Min: 0, 1st Qu.: 24, Median: 45, Mean: 42.5, 3rd Qu.: 62, Max: 100
track_album_id	Character	Album identifier
track_album_name	Character	Album name
playlist_name	Character	Name of the playlist the track belongs to
playlist_id	Character	Playlist identifier

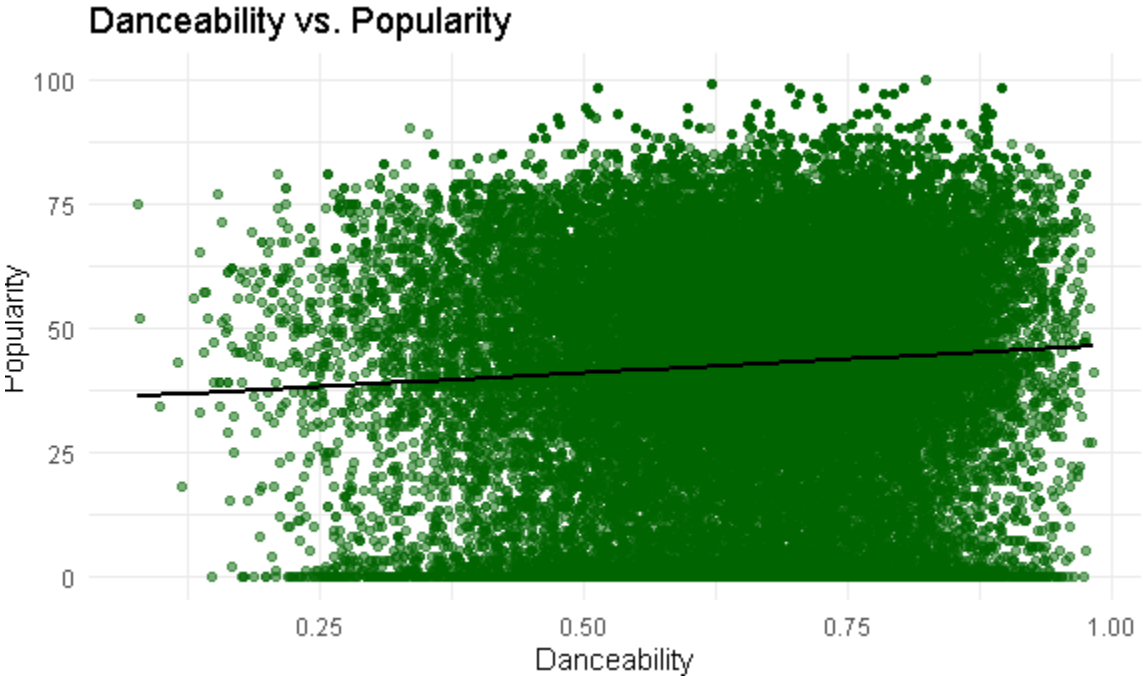
Variable	Type	Summary
playlist_genre	Character	Genre category of the playlist
playlist_subgenre	Character	Subgenre classification of the playlist
danceability	Numeric	Min: 0.077, Median: 0.672, Max: 0.983, Mean: 0.655
energy	Numeric	Min: 0.0002, Median: 0.721, Max: 1.0, Mean: 0.699
key	Integer	Range: 0–11, Mean: 5.37
loudness	Numeric (dB)	Min: -46.45, Median: -6.17, Max: -0.05, Mean: -6.72
mode	Binary	0 = Minor, 1 = Major, Mean: 0.57
speechiness	Numeric	Min: 0.022, Median: 0.063, Max: 0.918, Mean: 0.107
acousticness	Numeric	Min: 0.0000014, Median: 0.0804, Max: 0.994, Mean: 0.175
instrumentalness	Numeric	Min: 0.00000, Median: 0.00000, Max: 0.994, Mean: 0.085
liveness	Numeric	Min: 0.0094, Median: 0.127, Max: 0.996, Mean: 0.190
valence	Numeric	Min: 0.00001, Median: 0.512, Max: 0.991, Mean: 0.511
tempo	Numeric	Min: 35.48 BPM, Median: 121.98, Max: 239.44, Mean: 120.89
duration_ms	Numeric	Min: 31,429 ms, Median: 216,002 ms, Max: 517,810 ms, Mean: 225,818 ms
duration_min	Numeric	Min: 0.52 min, Median: 3.6 min, Max: 8.63 min, Mean: 3.76 min
album_release_year	Integer	Min: 1957, Median: 2016, Max: 2020, Mean: 2011

Exploratory Data Analysis

To better understand what makes certain songs more popular, I explored the dataset with eight key questions. Each question was designed to uncover trends or relationships from the raw data. I grouped, summarized, and visualized the data to extract deeper insights. Below, each question is followed by the insight it revealed and plot that supports the finding.

1. Do more popular songs tend to have higher danceability scores?

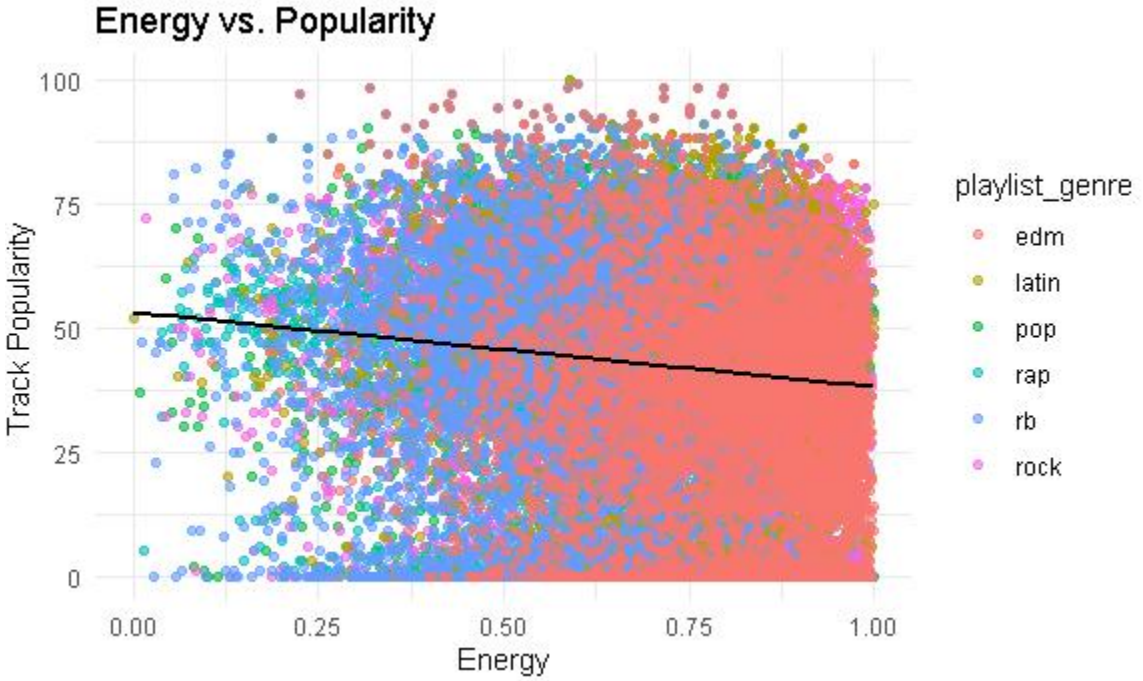
I plotted danceability onto a graph compared their popularity. I found that there was a slight upward trend with popularity and danceability increased, suggesting that danceability could be an influence on a song's popularity.



2. What energy levels are associated with the most popular songs?

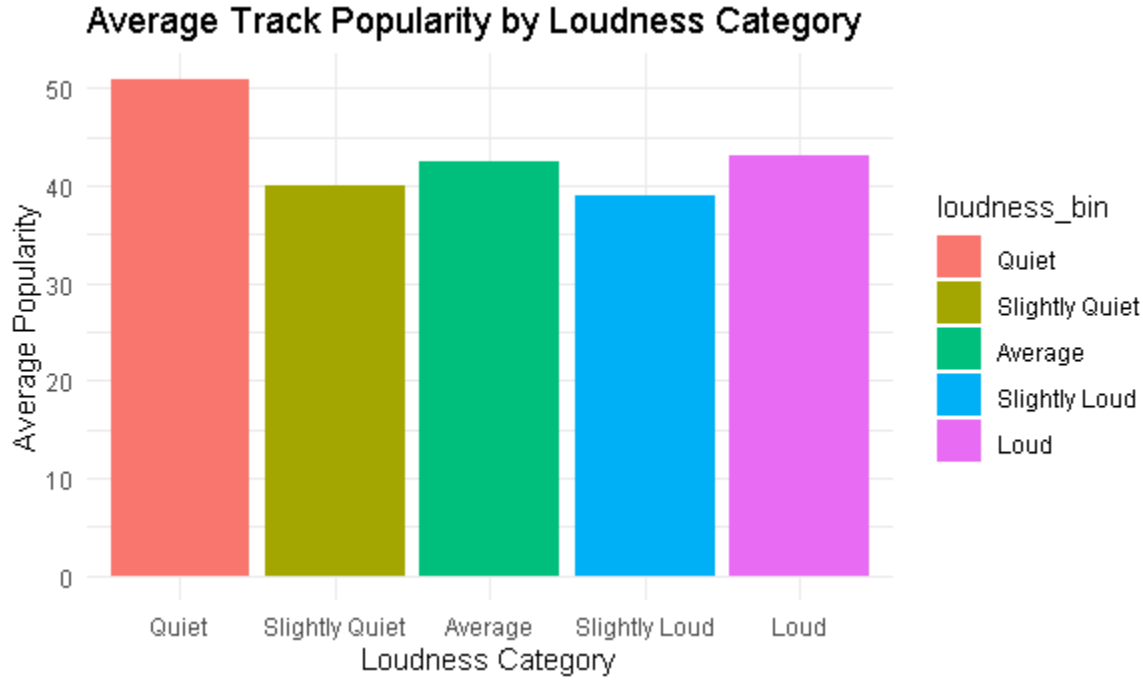
After analyzing the distribution of energy across popularity, it became evident that songs with higher energy are less likely to be popular. The correlation shows a downward trend in

popularity as energy rises showing energy is a factor in a song's popularity.



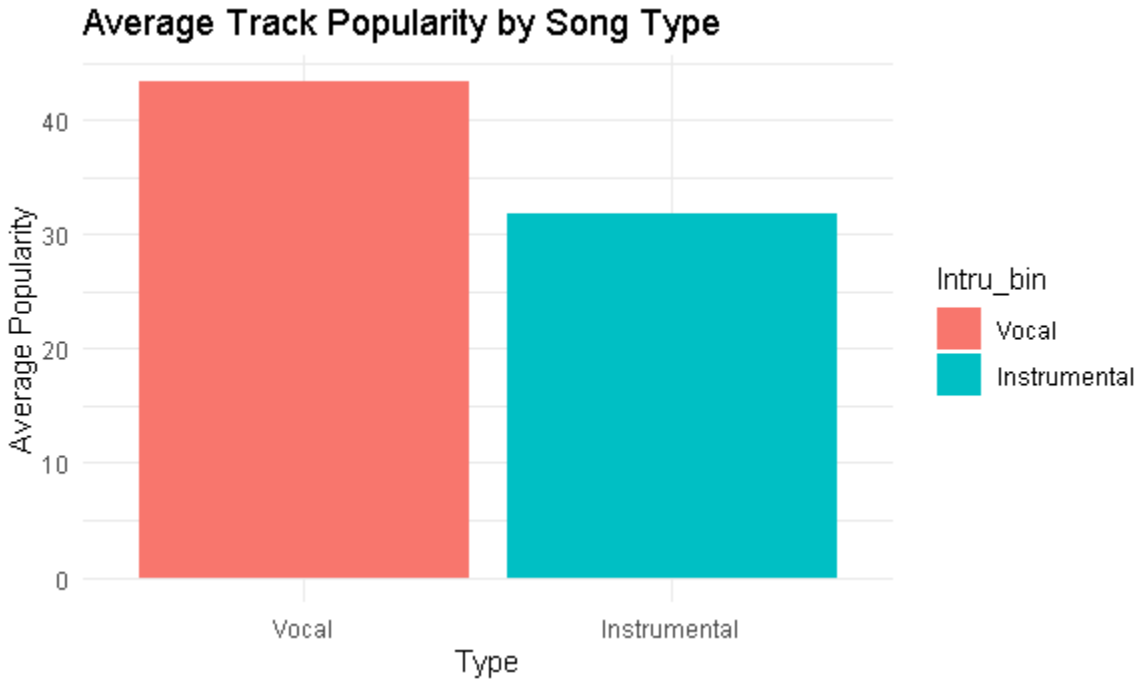
3. Is there a correlation between loudness and popularity?

I plotted loudness brackets against popularity and noticed everything is close with quite songs being slightly higher rated in popularity. This makes sense as higher energy songs ranked lower in popularity. This shows that loudness is a factor in a song's popularity.



4. How does valence (musical positivity) influence popularity?

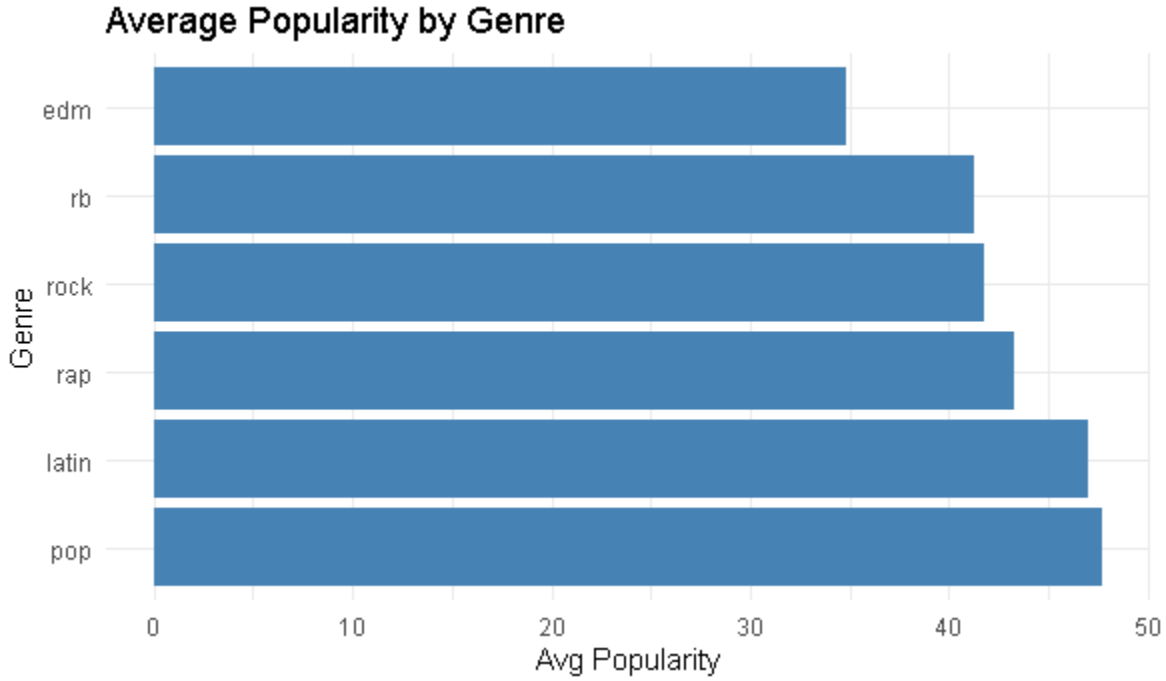
By categorizing instrumental only tracks, I observed that vocal songs on average had a higher popularity than songs that were instrumental only.



5. Which genres dominate the most popular tracks?

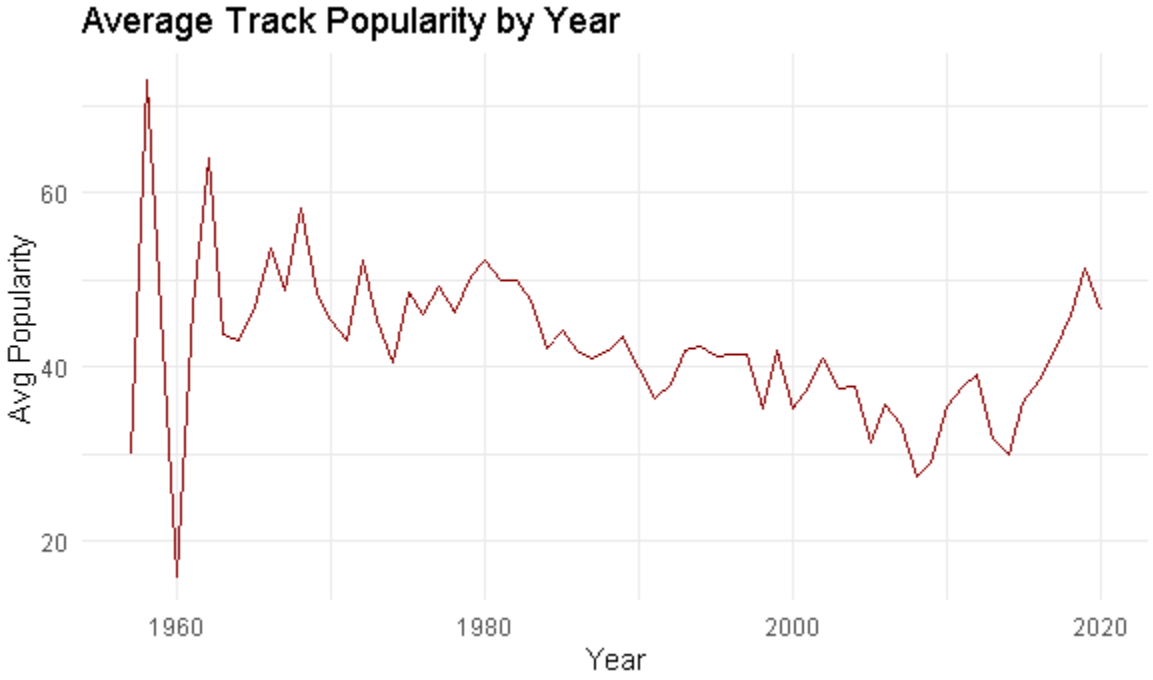
I created a bar chart of top songs by playlist genre. Pop and Latin showed the highest

popularity among the genres with average popularity just under 50.



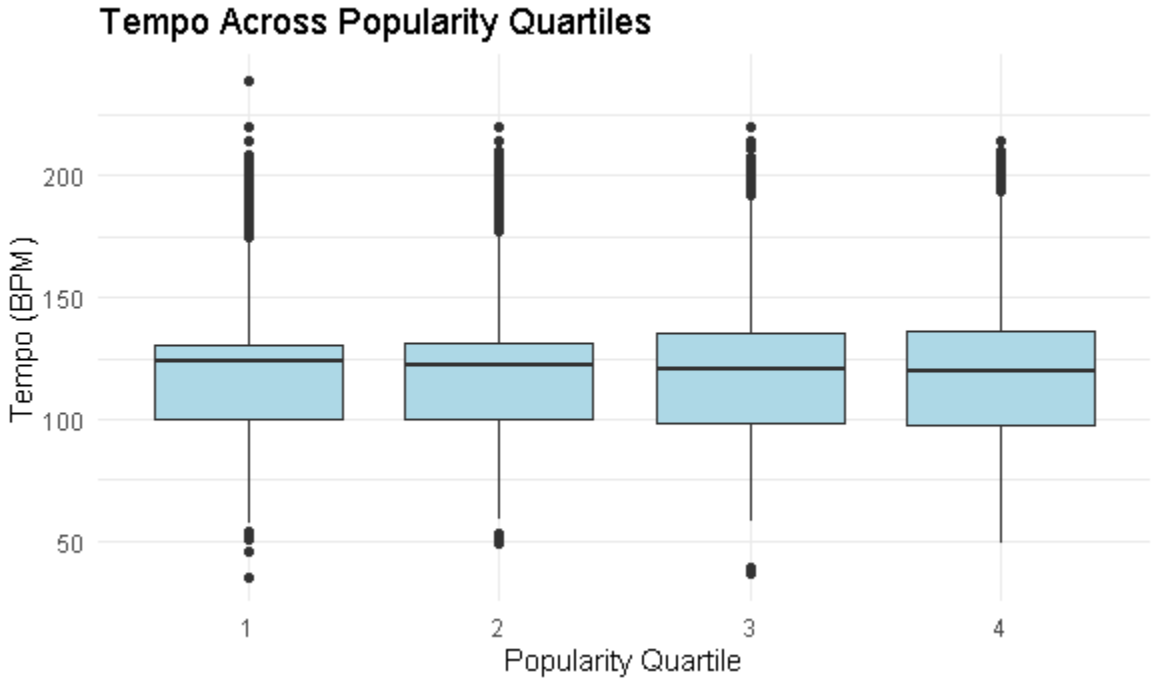
6. How has the popularity of songs evolved over the years?

By aggregating average track popularity by year, I found that there as a slight downward trend until newer songs, especially post-2015, tend to dominate in popularity, compared to recent years, likely due to streaming platform trends.



7. What is the relationship between tempo and popularity?

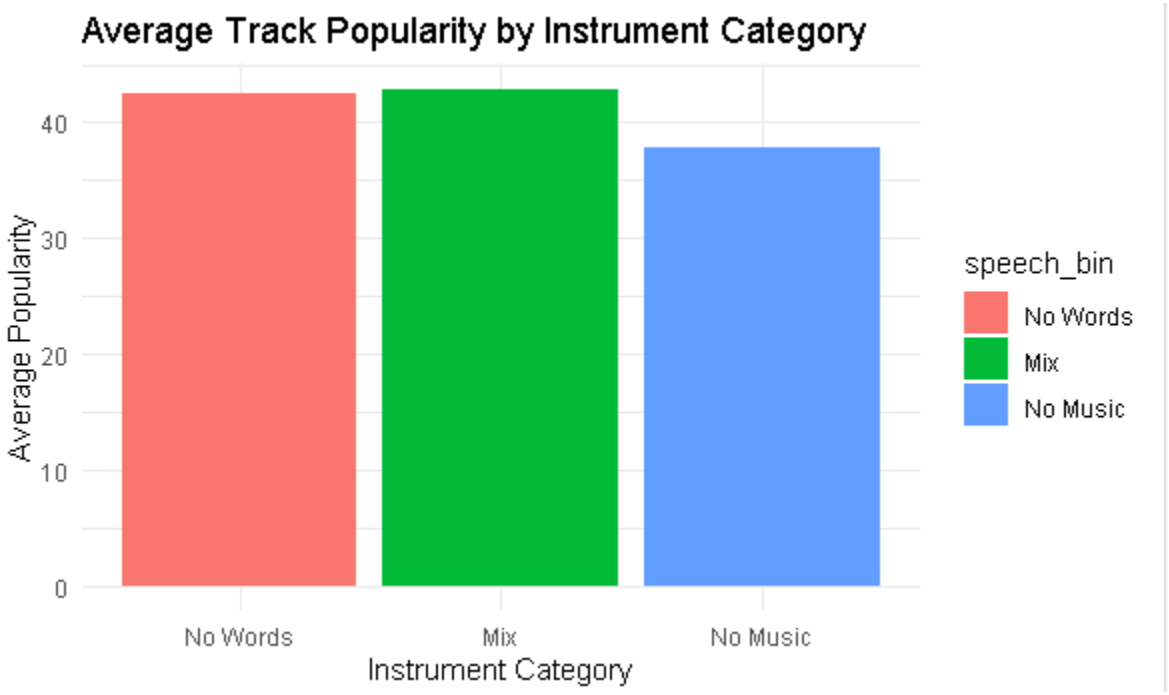
I analyzed the tempo of popular vs. less popular songs. Across the board it looks similar and seems tempo is not a factor in popularity.



8. Is there a difference in speechiness between more and less popular songs?

I grouped songs into speechiness labels and calculated average popularity. Popular tracks were those with lower speechiness on average, which aligns with Spotify's musical focus

rather than spoken word content.



Each of these questions helped transform raw variables into meaningful insights. Through this process, we revealed subtle relationships and trends that contribute to a song's popularity, information that could be useful for artists, producers, and marketers alike.

Summary

The problem addressed in this analysis is identifying the factors that influence song popularity. By posing eight key questions around attributes such as danceability, energy, loudness, and genre, the goal was to uncover trends and relationships that could explain why certain songs are more popular. The analysis revealed several insights: danceability showed a slight upward trend with popularity suggesting it could be an influencing factor, energy was inversely correlated with popularity with higher energy tracks generally less popular, louder songs tended to have slightly lower popularity, vocal tracks were more popular than instrumental ones, and genres like Pop and Latin dominated in popularity. There was a slight decline in popularity until post-2015, where a resurgence in popularity was observed, likely due to streaming trends. Tempo did not significantly affect popularity. Popular songs had lower speechiness, aligning with Spotify's preference for musical over spoken content. These findings help clarify what attributes contribute to a song's success and provide valuable insights for the music industry.

Presentation Link:
https://drive.google.com/file/d/1ivx7intEQ_CAKTABxpK042pRiCcw7dJ0/view?usp=drive_link