Maximus Wu
Professor Koehler
Data Bootcamp
21 March 2025

Midterm Project Writeup

Project Link:
https://colab.research.google.com/drive/1Cfqx4lxWEajYDxBok1J0PJ8dE6exjRF4?usp=sharing

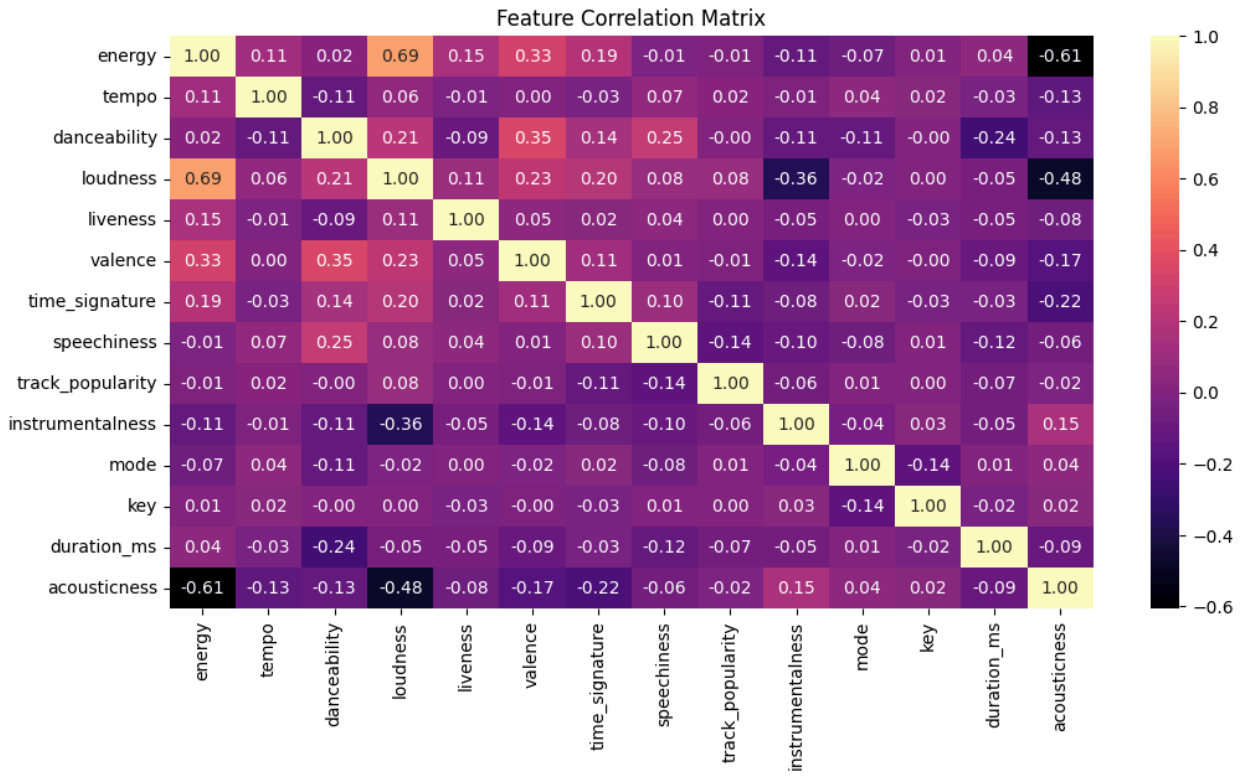
Introduction and Objectives of the Project:
This project aims to explore the relationship between various musical features and track popularity. Using a dataset containing information on tempo, energy, danceability, and other audio characteristics, I wanted to:

1. Identify what factors most to making a 'hit' song
2. Evaluate the evolution of popular music, and see what characteristics of music were changing over time
3. Conduct some statistical analysis on music streaming patterns to see if there were some traits that were necessary or at least very correlated to a music's popularity.

Methodology & Data Collection:
The data that I am using for the project was taken from a database on Kaggle, courtesy of Solomon Ameh. According to the description, the data was collected via Spotify's APIs to collect different data about some of the most popular songs. I've linked the original database here.

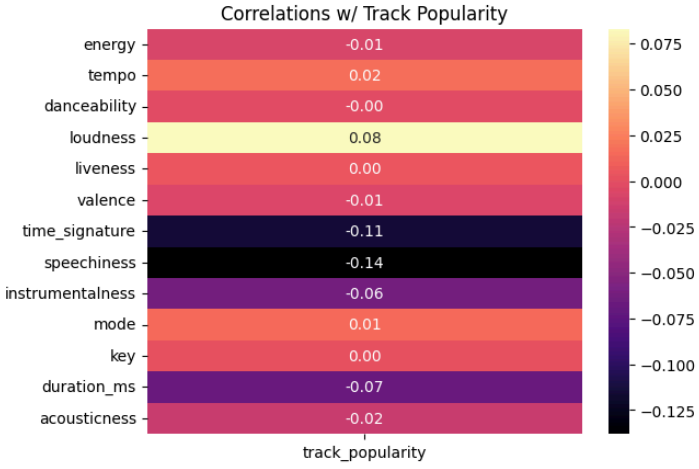
Objective 1: Identify what Factors Most to Making a 'Hit' Song

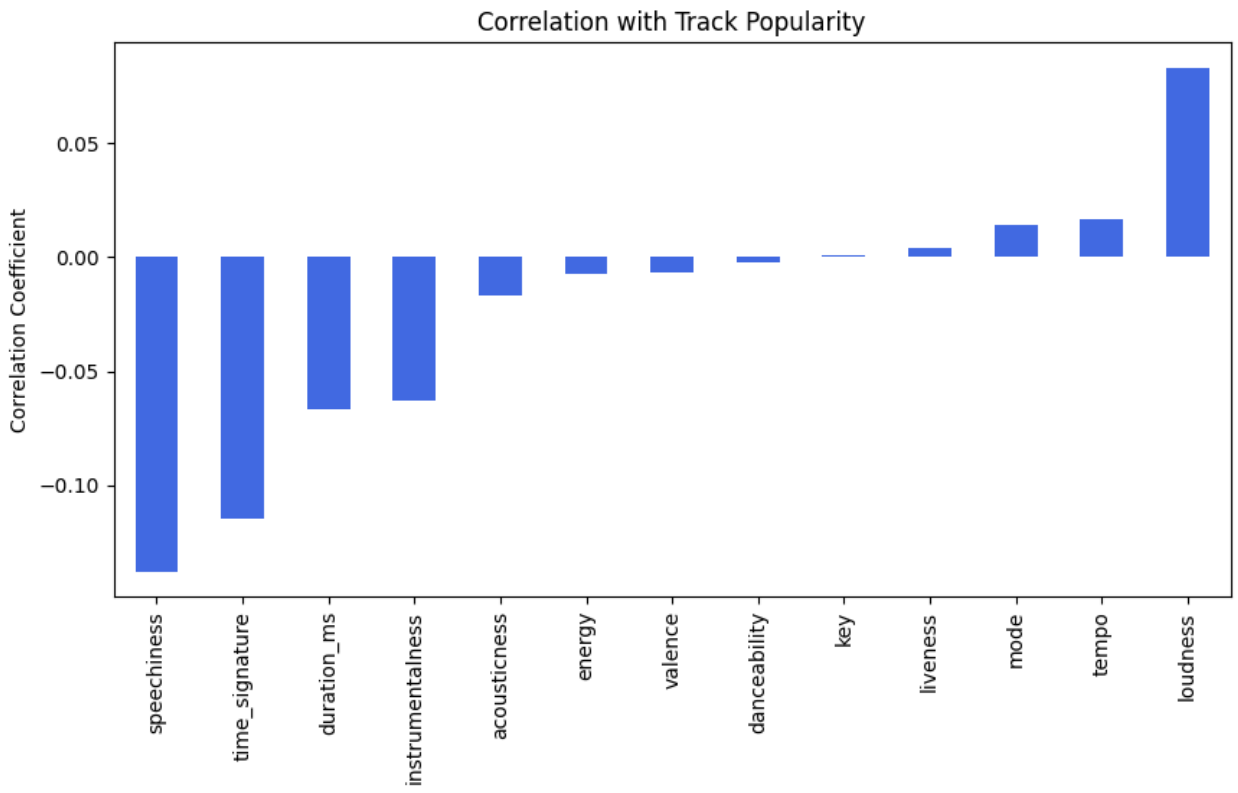
Exploratory Data Analysis (EDA)
To kick off the project, I wanted to do some initial data analysis and get a good understanding of how certain variables correlated to others. I figured the best way to go about this was creating a heatmap and seeing all of the correlations between quantitative variables, shown below.



As we can see, there are not a lot of variables that correlate heavily with each other. The only somewhat strong correlation that I can see is that loudness correlates with energy, which is relatively self-explanatory. Weaker correlations include danceability and valence, along with valance and energy, which again seems relatively self-explanatory as valence is defined as 'musical positivity' conveyed by a track.

From this massive heatmap, I wanted to then hone in specifically on one column, the track popularity, addressing one of our key questions as that would hopefully help us get a better understanding of what factors correlate with popularity.



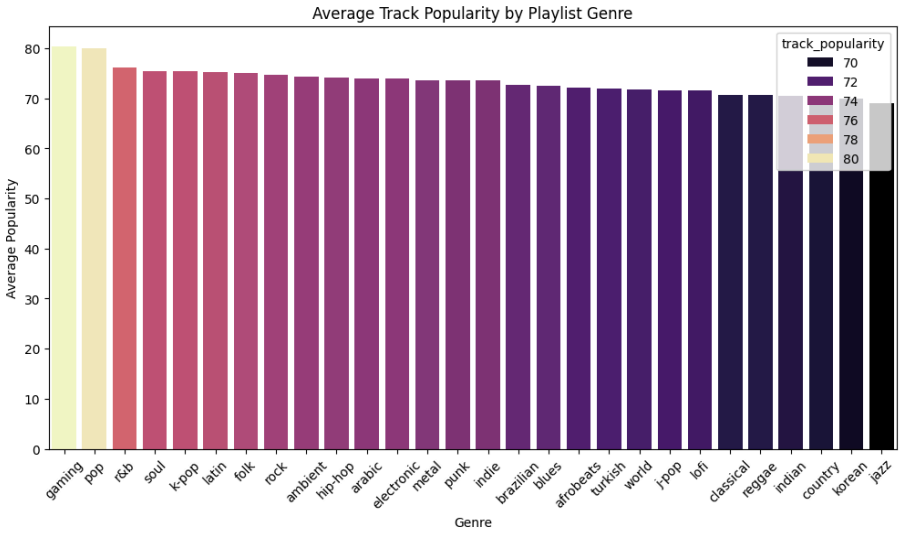


From these results, we see the:

1. The largest correlation with track popularity today is 'loudness'
    a. For me, this actually makes a good amount of sense since most songs that tend to be pretty popular are ‘poppy’ and upbeat, and therefore louder than more somber songs.
2. The largest Inverse Correlation w/ popularity is 'speechless', therefore less wordy songs seem to actually be more popular
    a. This was relatively shocking to see since a lot of today’s most popular songs are hip-hop and rap, which tend to be more wordy. Especially during the early 2000s, when artists like Eminem were extremely popular, it would seem contradictory to say that speechiness has an inverse relationship with popularity. It could be possible that ‘speechy’ rapping styles have fallen out of popularity relative to more melodic flows like that of Drake/Kendrick Lamar.
3. Additionally, smaller factors like duration and time signature play a role in inversely correlating with the most popular songs on Spotify currently
    a. These make a lot of sense as well since most ‘pop’ hits that we see are typically around 2-4 minutes long. Any longer than that, and it becomes repetitive, and any shorter tends to not be popular either as it’s too short. Later on in this project, I built out a polynomial graph that plotted popularity with track duration, and it seems like the

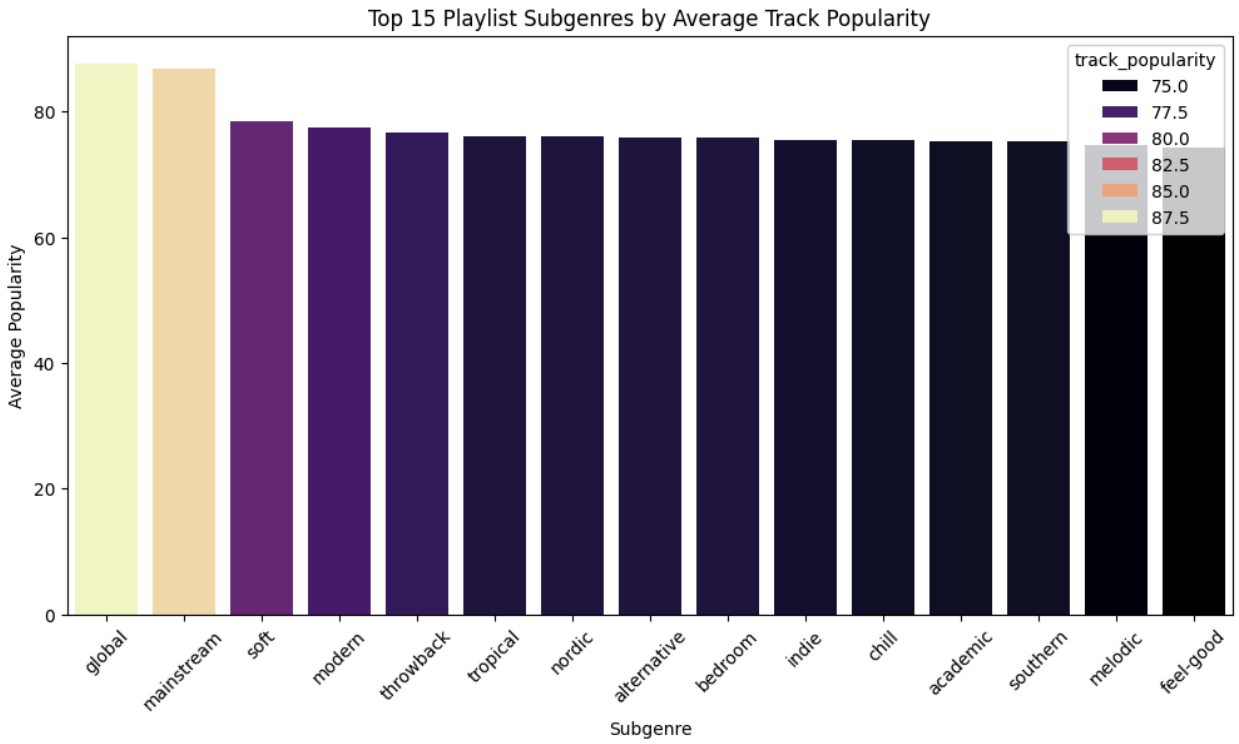
most popular ones are within the 150 to 200-second length which aligns well with my prediction from earlier.

Following up, I wanted to then move toward the categorical variables to see if there were any major correlations between factors like genre/artist with popularity. Starting with genre, I plotted average track popularity by the different genres:



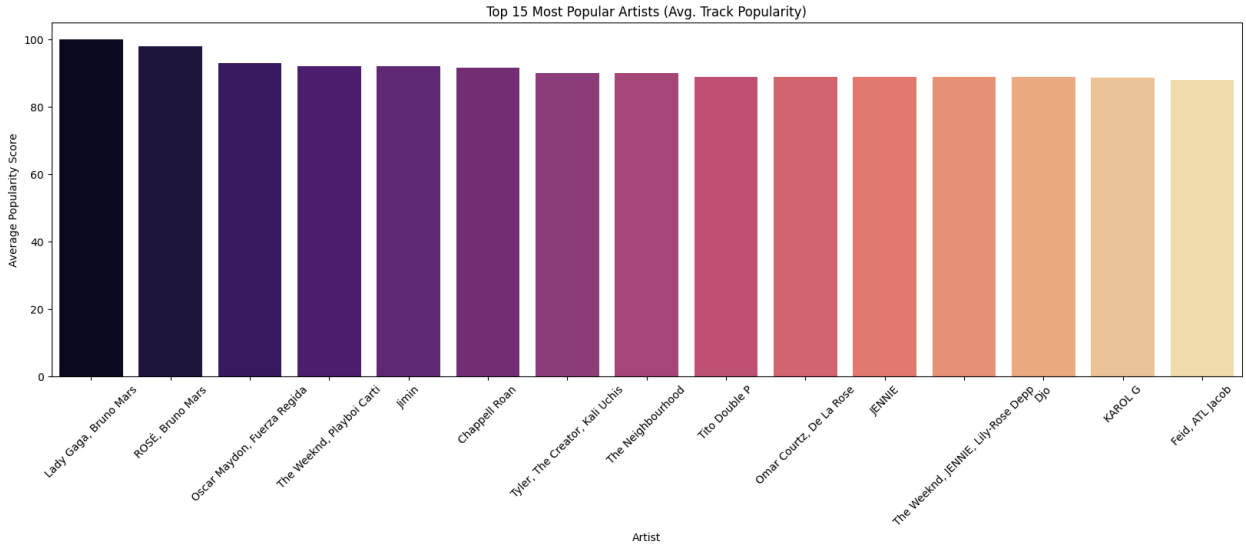
I'd say that most of the results make sense with what I expected, with pop and R&B being at the top, however, I was definitely not expecting 'gaming' to be close to the top, let alone taking the number one spot. I'm still a little confused by this result, but I assume that it's up there because of how niche it is, and therefore has a very small but strong followerbase which could incease the average popularity of songs in that category.

Because I wasn't satisfied with those results, I wanted to take a deeper look into subgenres to see which ones were most correlated with popularity:

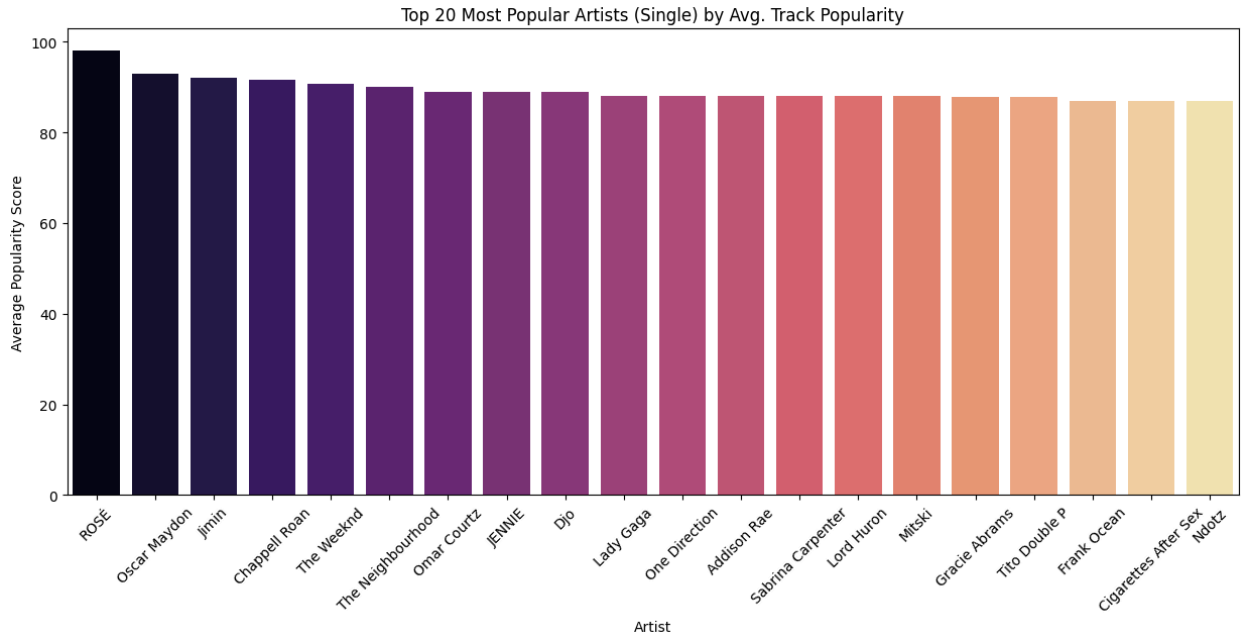


Taking the top 15 most popular subgenres, I'd say that the top 15 are definitely in line with what I expected, with global, mainstream, soft, and modern all being in the top 5.

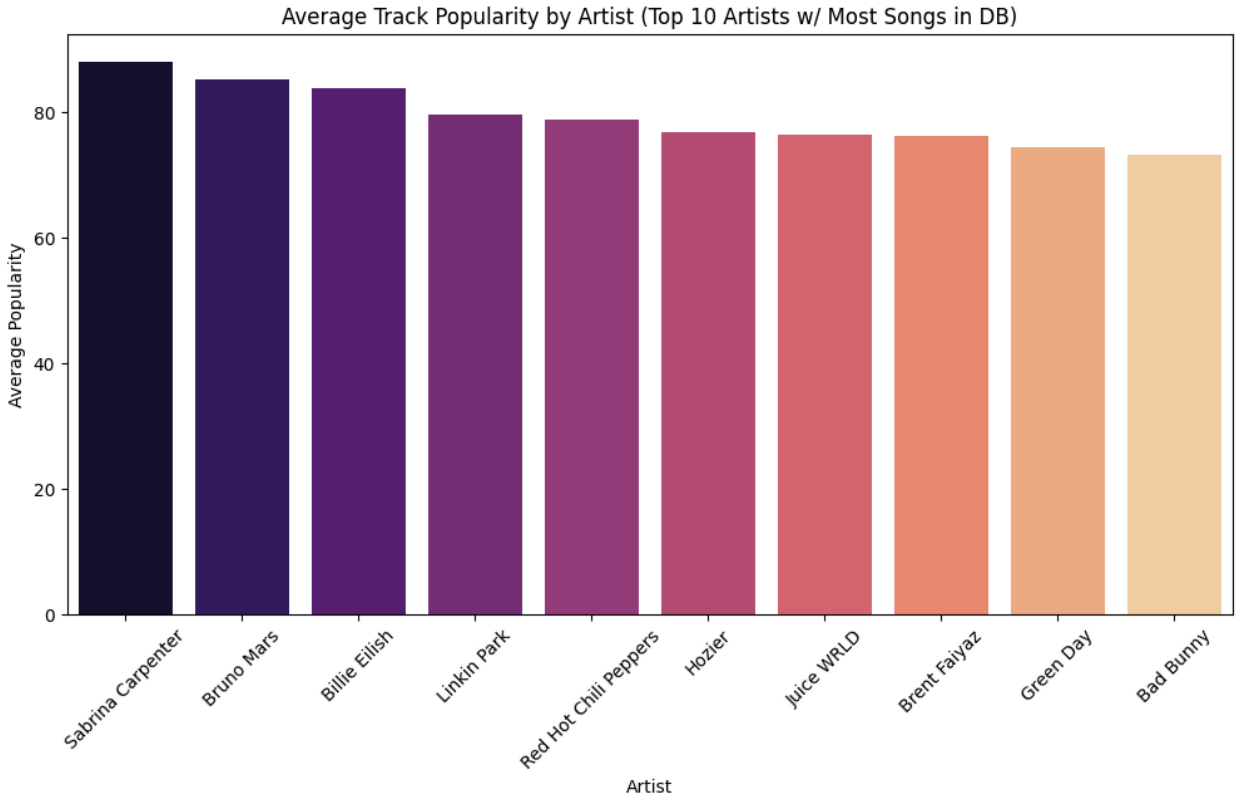
Secondly, I took a look at the top 15 artists that had the highest average popularity score



However, from these results, we see that it includes multiple artists in some columns, which is definitely not something that we are looking for. I hypothesize that it is doing this because there are a handful of super popular tracks that have certain combos of artists, (ie the top result being Lady Gaga & Bruno Mars's "Die with a Smile").



Lastly, I wanted to evaluate whether artists who created more music also correlated with the top-performing artists in the dataset. Below I listed the top 10 artists by song count in the database.

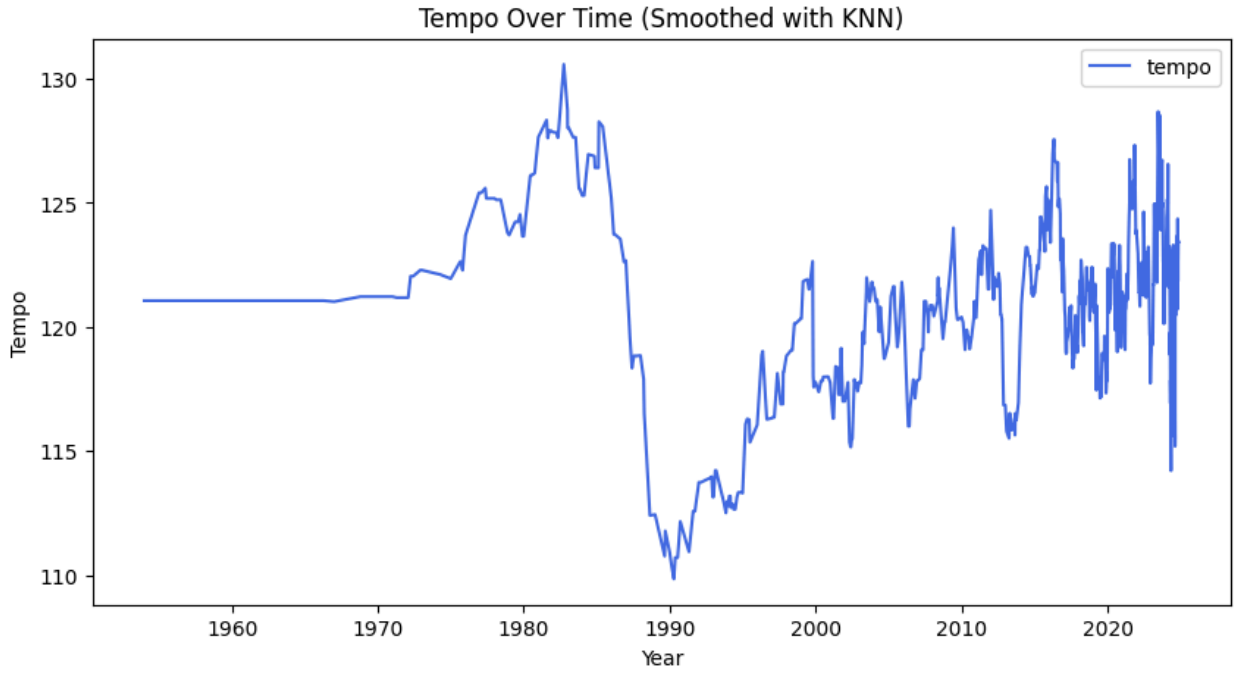


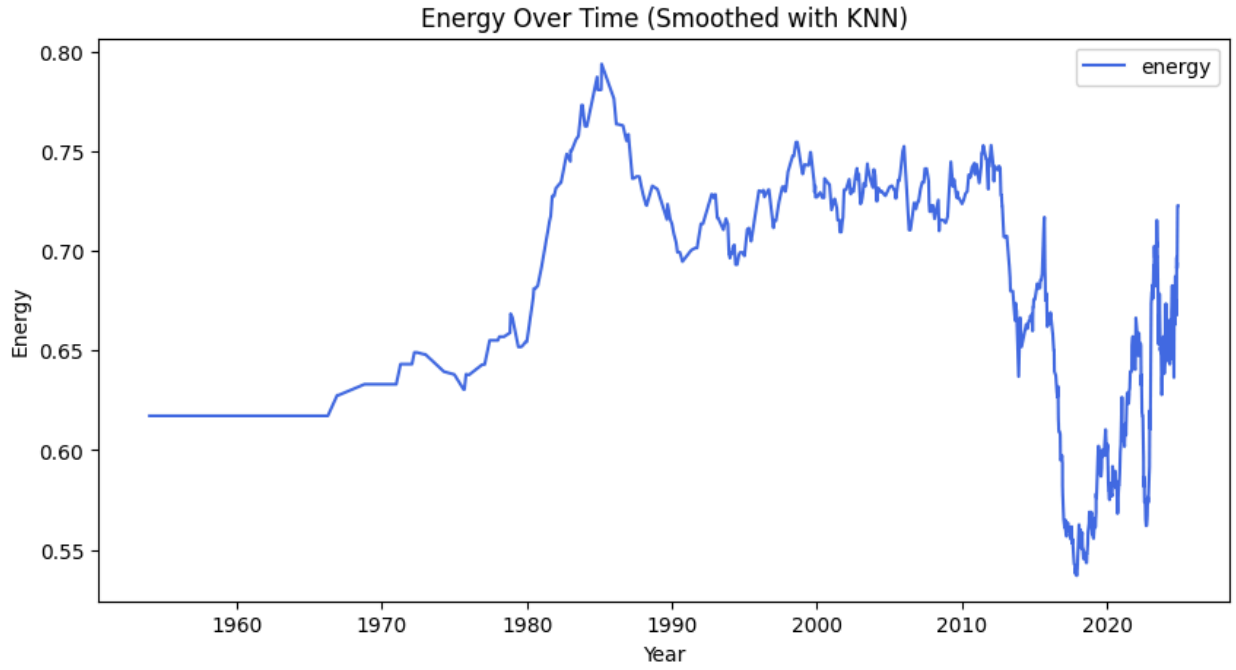
There is only minimal overlap between the two, with just Sabrina Carpenter being on both. Similar to the hypothesis from earlier, this makes sense because as there are more songs (test points), the average tends to converge to its true value.

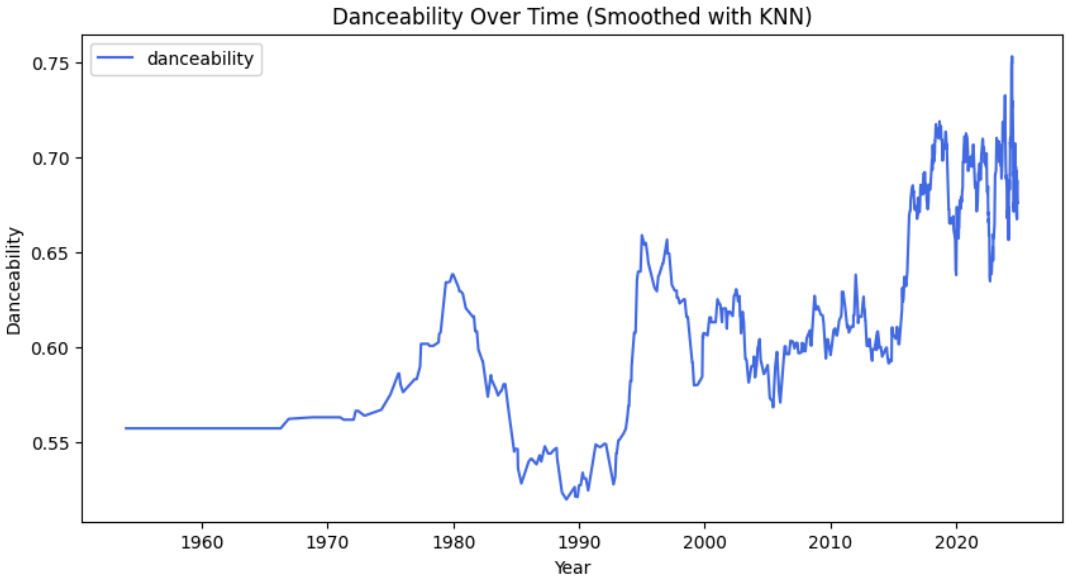
Objective 2:Evaluate the Evolution of Popular Music

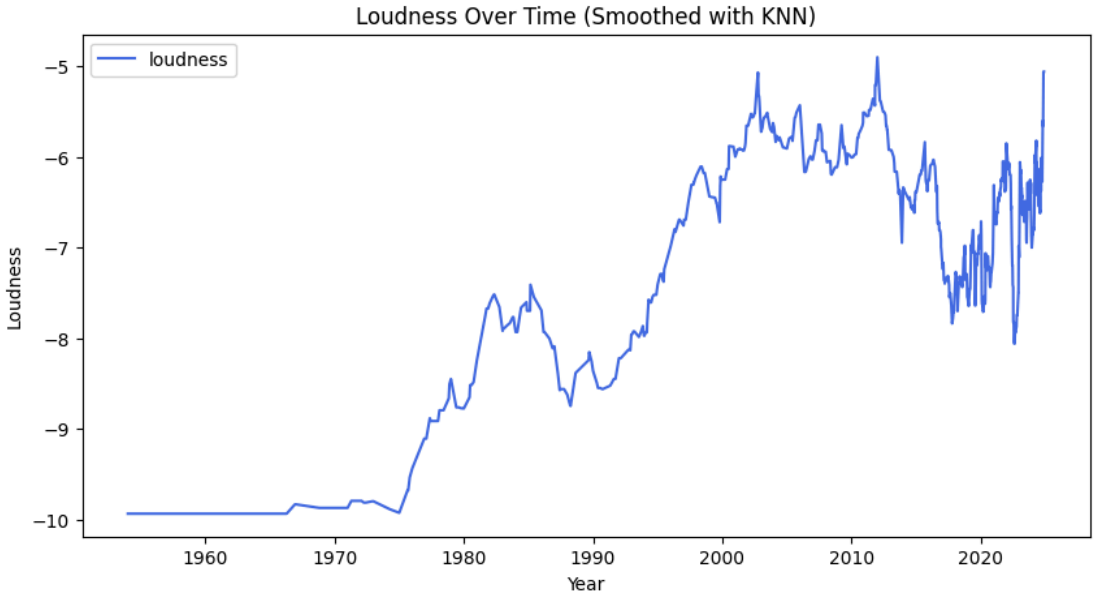
Time Series Analysis

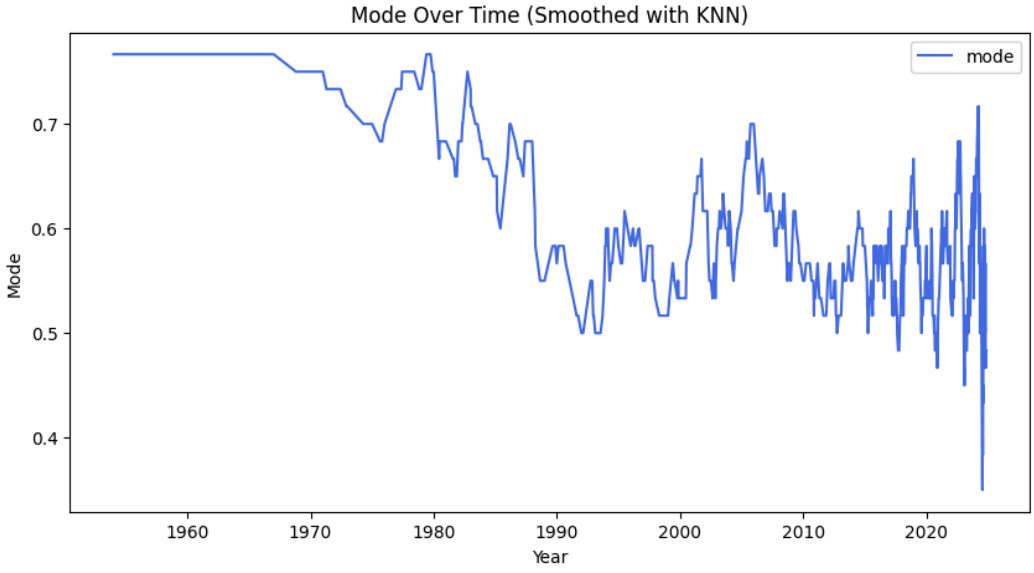
For this section, I created a series of time graphs that tracked how various characteristics of songs evolved over time. For this experiment, I chose 5 quantitative factors: Tempo, Energy, Danceability, Loudness, and Mode. To create this graph, I used a KNN model to smooth out the lines, using 60 neighbors and tracked how it changed based on the release date. I found that 60 neighbors were the best number after playing around a little with the graphs, as any less would look too volatile and any more seemed unnaturally smoothed. The resulting 5 graphs are all pasted below.


Tempo Over Time (Smoothed with KNN)
tempo
Tempo
130
125
120
115
110
1960
1970
1980
1990
2000
2010
2020
Year


Energy Over Time (Smoothed with KNN)
energy
Energy
0.80
0.75
0.70
0.65
0.60
0.55
1960
1970
1980
1990
2000
2010
2020
Year


Danceability Over Time (Smoothed with KNN)
danceability
Danceability
0.75
0.70
0.65
0.60
0.55
1960
1970
1980
1990
2000
2010
2020
Year


Loudness Over Time (Smoothed with KNN)
loudness
Loudness
−5
−6
−7
−8
−9
−10
1960
1970
1980
1990
2000
2010
2020
Year


Mode Over Time (Smoothed with KNN)
mode
Mode
0.7
0.6
0.5
0.4
1960
1970
1980
1990
2000
2010
2020
Year

Results from trend analysis:

1. Tempo in popular music dropped significantly in the 80s but made a steady rise ever since the early 90s to the modern day
2. Energy in top songs took a major drop in the 2010s, however, it again made a large comeback in the late 10s/early 20s
3. Since the start of the 90s, danceability has been an increasingly large indicator of popular music
4. Likewise, we see that loudness has also increased a lot over time, with major increases in the 90s decade, but dropped a little during the 2010s, with a massive comeback in the 20s.
5. Lastly, mode (I believe 1 meaning major and 0 meaning minor) has seen a massive shift towards minor scales in the last few decades.

Objective 3: Statistical & Modeling Work on Dataset

Linear & Polynomial Regression Models

I attempted to use 2 predictive models to explain track popularity: linear & polynomial regression.

```
Linear Regression CV Scores: [-0.01290582  0.01973917 -0.33841649 -0.32742642 -0.43085914]
Average CV Score: -0.2179737416694884
Polynomial Regression CV Scores: [-0.0246888   0.0034625  -0.30408848 -0.44568855 -0.81043379]
Average Polynomial CV Score: -0.3162874222758679
```

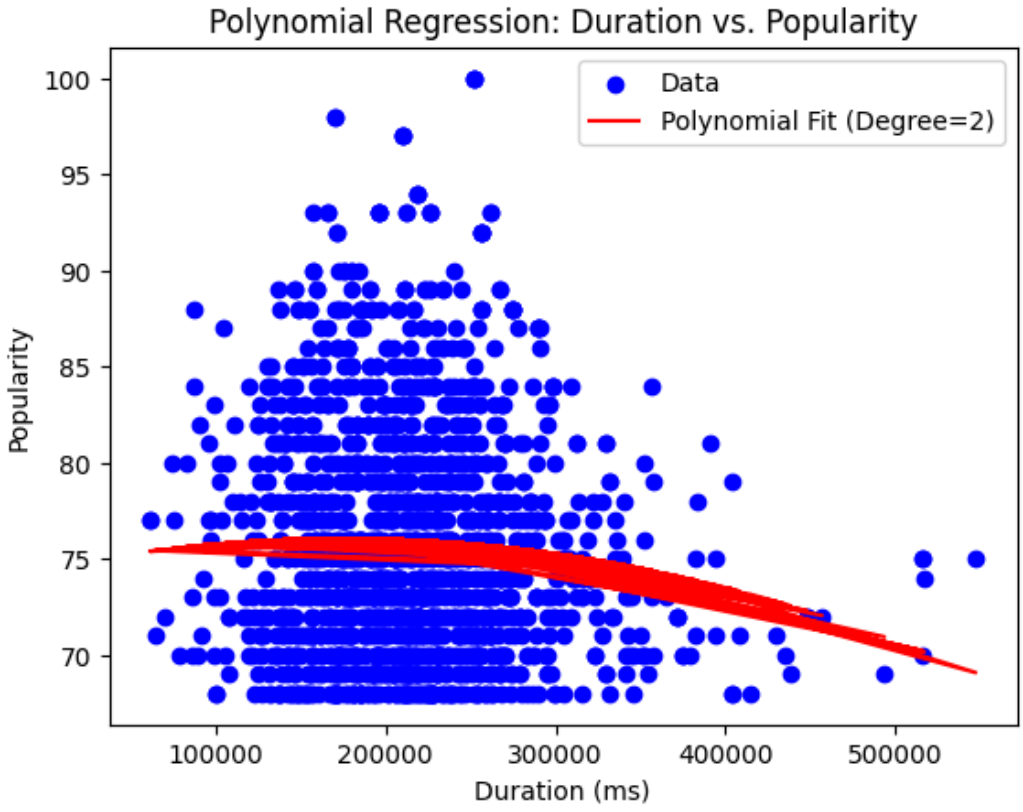
Clearly from the results, both performed poorly with linear regression having an average cross-validation score of -0.218. The negative score indicates to us that the model did worse than simply predicting the average popularity for all songs. Ultimately, this linear regression model suggests that the relationship between quantitative features and popularity is relatively weak/nonlinear. However, I then proceeded to make a polynomial model which did even worse than the linear one, with an average polynomial CV score of -0.316. I assume that the model here overfit the training data and therefore failed to generalize well with new data. As a whole, it seems that there is no polynomial relationship that song popularity follows, which I believe to actually be good for the music industry as a whole, as there isn't a defined "recipe" to follow to create a hit song.

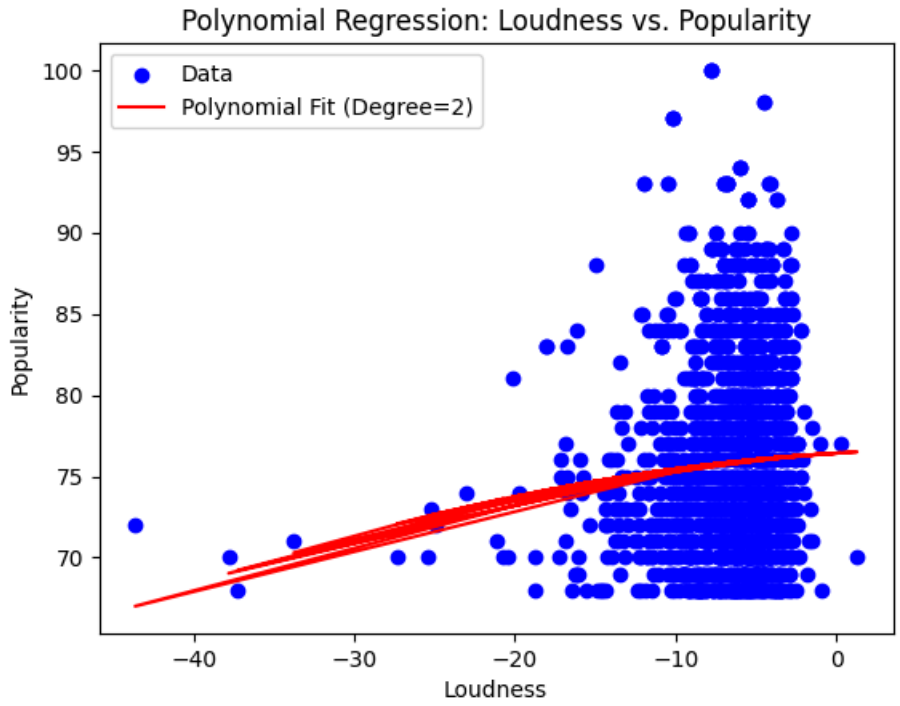
Additionally, I conducted a KNN Regressor, which smoothed some factor trends but it still didn't perform that well since the average CV score was still negative. Overall, none of these models proved to be very effective at predicting popularity.

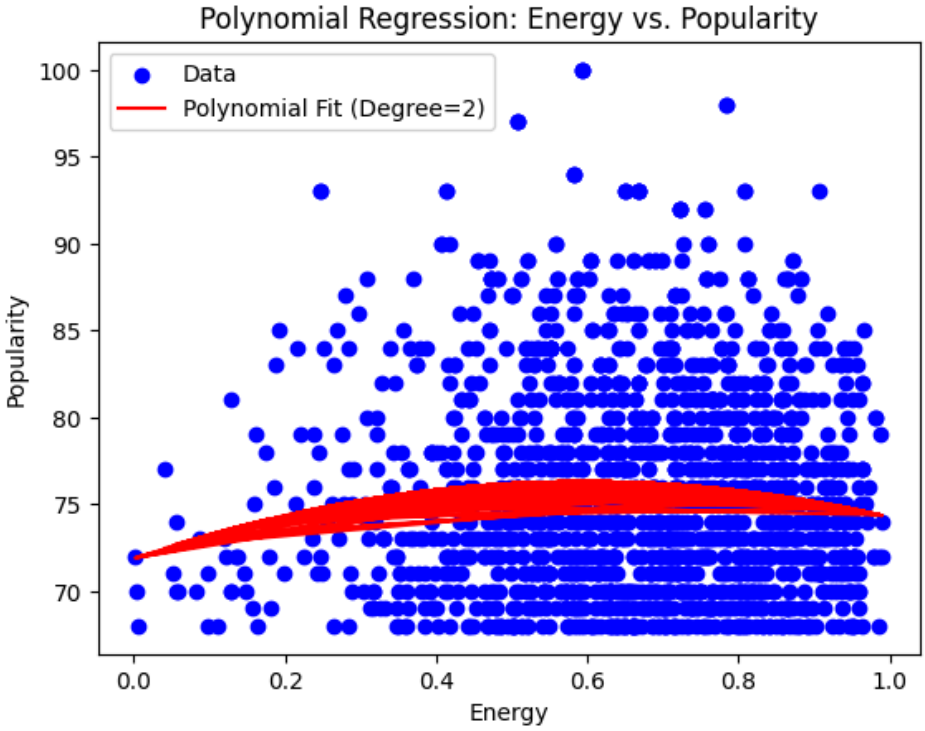
```
KNN Regressor CV Scores: [-0.03815169 -0.04533732 -0.01121533 -0.05481018  0.03744578]
Average KNN CV Score: -0.022413748442328463
```

Lastly, I conducted some statistical tests to see whether there was a statistically significant factor that differed between normally popular and the most popular songs. I split the dataset into two, popular and normally popular based on whether a song's popularity was greater/less than the median. From there I iterated through the various factors and saw which factors had a statistically significant difference between the most/normal popular songs. In the end, we only saw that 'duration' had a somewhat noticeable effect, with a t-stat of -0.95 (less than one standard deviation away from the mean) and a p-value of 0.34 (which is above our threshold of 0.05). Ultimately, it seems like there is no sole predictor of a song's success.

Lastly, I wanted to do some additional plotting of factors w/ popularity. The three I thought were most interesting were duration, loudness, and energy. Below are the polynomial fit trends.




Polynomial Regression: Loudness vs. Popularity
Data
Polynomial Fit (Degree=2)
Popularity
Loudness
100
95
90
85
80
75
70
−40
−30
−20
−10
0


Polynomial Regression: Energy vs. Popularity
Data
Polynomial Fit (Degree=2)
Popularity
Energy
100
95
90
85
80
75
70
0.0
0.2
0.4
0.6
0.8
1.0