Spotify Music 1921-2023 Analysis

Problem Statement

Evolution is a defining life theme. Regardless of the measure... people, technology, fashion, or even music, these aspects of life evolve under the pressure and impact of both one another and from the societies behind them from generation to generation. Each generation differs from the next and that is what is both telling and interesting. Why were certain measures defined in the ways they were to one generation, but the same measures look and feel much different to the next generation? In this project, I will explore the underlying music trends spanning 1921-2023, and I will leverage the correlations amongst the data to help my clients better understand the ever-changing music industry.

Data Wrangling

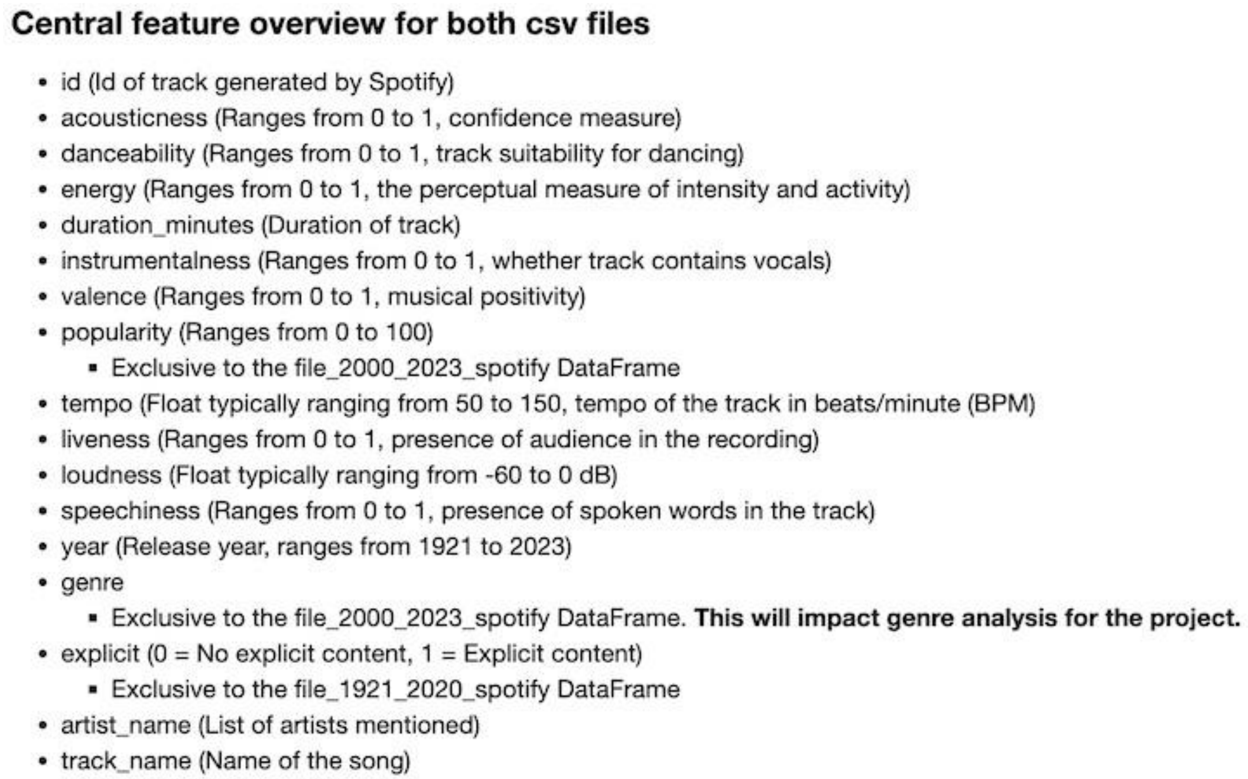
I leveraged two datasets, and both are quite clean as is, so I did not need to apply drastic changes. Though, touch-up was needed to enhance the interpretability for my clients.

Changes made include:

- 15 values or 0.13% of the artist_name field were missing. Filled with value "other" in Dataset 1.
- 1 value or 0.01% of the track_name field was missing. Filled with value "other” in Dataset 1.
- Converted track duration from milliseconds to minutes for better interpretability in both Datasets.
- Renamed fields for better interpretability in both Datasets.
- Deleted duplicate field in Dataset 2.
- Stripped brackets and quotes in artist_name field in Dataset 2.

Exploratory Data Analysis (EDA)

Digging into the data, many trends and findings were uncovered. They are articulated below. First, it is vital to understand the features to then understand their respective correlations and trends amongst one another.



Central feature overview for both csv files

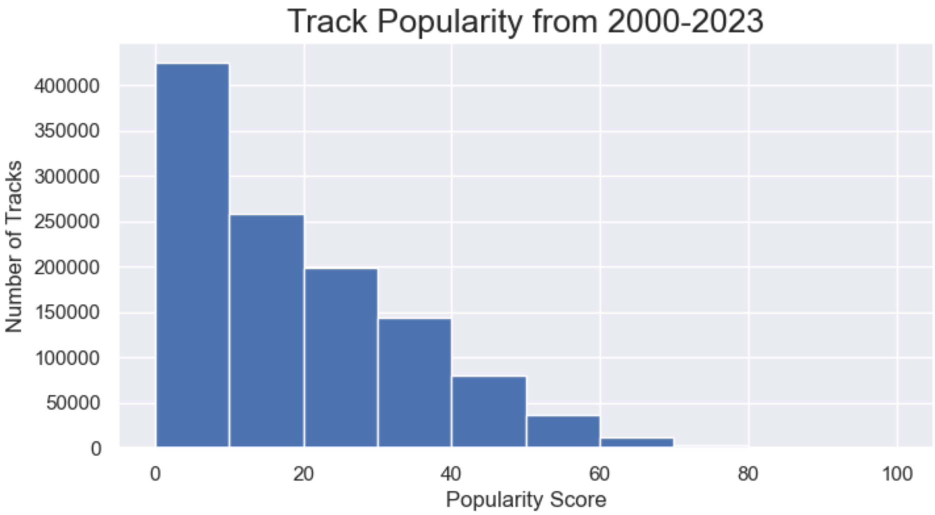
- id (Id of track generated by Spotify)
- acousticness (Ranges from 0 to 1, confidence measure)
- danceability (Ranges from 0 to 1, track suitability for dancing)
- energy (Ranges from 0 to 1, the perceptual measure of intensity and activity)
- duration_minutes (Duration of track)
- instrumentalness (Ranges from 0 to 1, whether track contains vocals)
- valence (Ranges from 0 to 1, musical positivity)
- popularity (Ranges from 0 to 100)
  - Exclusive to the file_2000_2023_spotify DataFrame
- tempo (Float typically ranging from 50 to 150, tempo of the track in beats/minute (BPM)
- liveness (Ranges from 0 to 1, presence of audience in the recording)
- loudness (Float typically ranging from -60 to 0 dB)
- speechiness (Ranges from 0 to 1, presence of spoken words in the track)
- year (Release year, ranges from 1921 to 2023)
- genre
  - Exclusive to the file_2000_2023_spotify DataFrame. **This will impact genre analysis for the project.**
- explicit (0 = No explicit content, 1 = Explicit content)
  - Exclusive to the file_1921_2020_spotify DataFrame
- artist_name (List of artists mentioned)
- track_name (Name of the song)

<u>Key Non-Feature Findings are Below:</u>

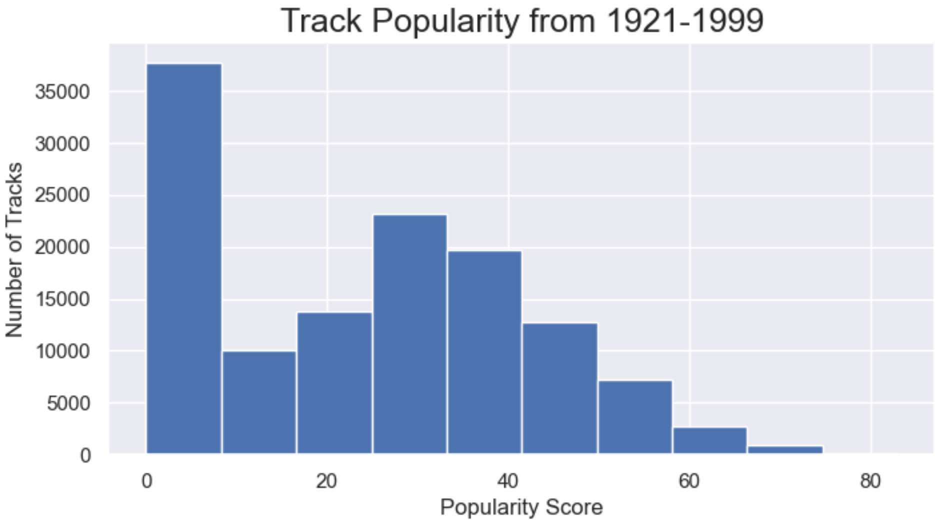
- There are 82 unique genres spanning 2000-2023
  - The 1921-1999 dataset (Dataset 2) does not have the "genre" feature, so I couldn't measure it.
- Songs have become noticeably less acoustic over time as the average value was around 0.7-1 from 1921-1964, then the range became 0.42-0.61 from 1965-1975, then the range hung around 0.30 or so from 1976-2023. Meaning, music has shifted to being louder/more energetic/more tempo based/etc. Perhaps this is due to the creation of more genres over time.
- Track duration also has wavered and has a pattern of starting out at 3.83 in 1921, declining to sub 3 minutes or so up until increasing to sub 4 minutes or so around the 1940s, it held strong and push sub 4.5 minutes or so for decades up until 2017 where it declined once again to the sub 4 minute or so range. Ironically, 2023's average track duration is 3.8 which is very close to 1921's value. See graphs below!
- Liveness has remained mostly unchanged, valence has not experienced much change either, and speechiness has remained mostly the same, though, it has experienced more outlier values across random years.
- As expected, energy has increased over time, and it has held strong in the sub 0.60-00.65 range from 1979-2023. And danceability has mostly remained the same which makes sense as music is generational hence, danceability is subjective and relative to its

era. Lastly, loudness has increased over time which likely is due to the advancement of technology, the creation process (ex. use of computers, studios, etc that previous generations did not have access to), generational appetites, and so on.

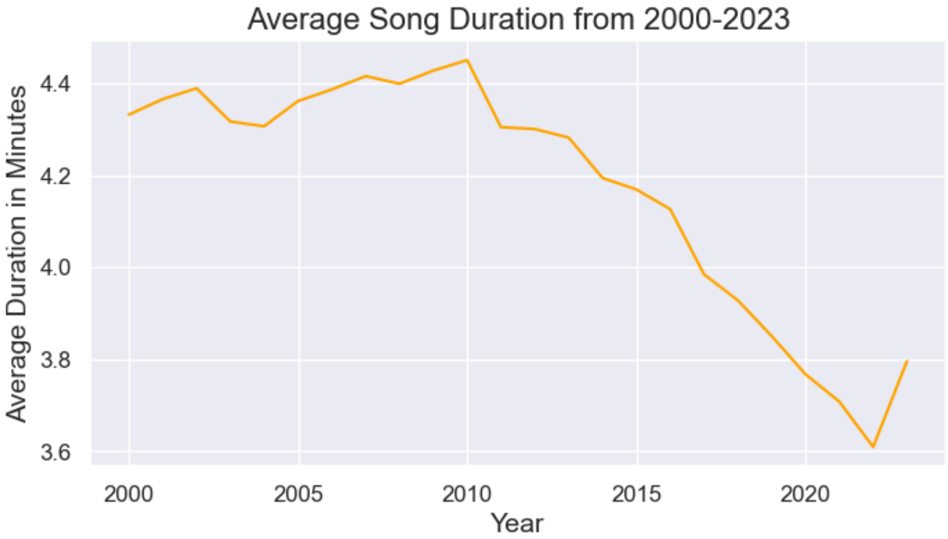
- Tempo has gradually increased over time and its range between the maximum and minimum values are much tighter than I anticipated (123.41-100.40).

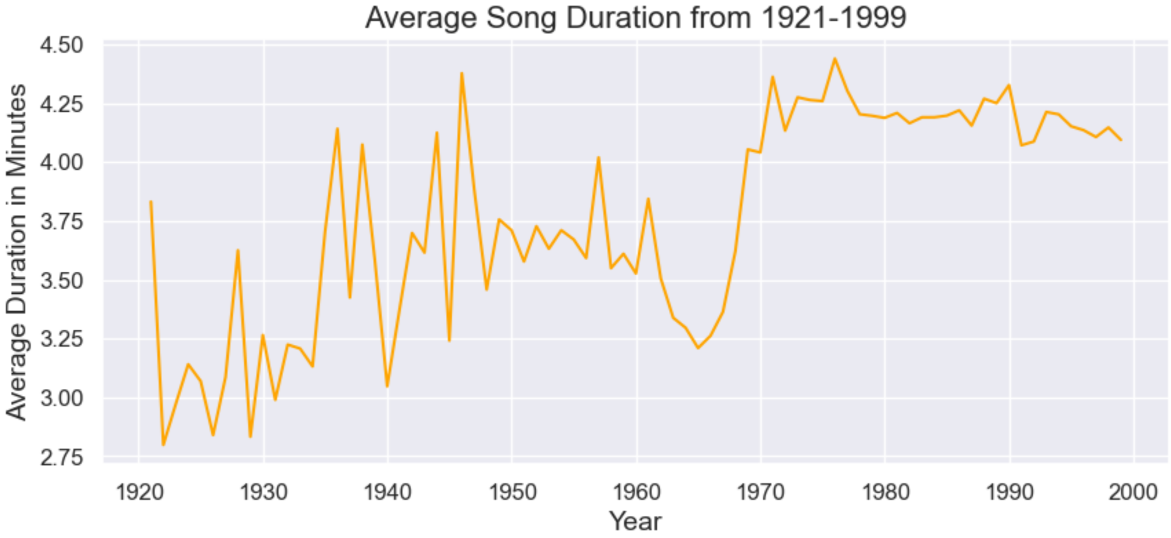


  - Over 400k tracks had a popularity score between 0-10. That is astounding and reflects close to 1/3 of the data alone. Inversely, only 15k tracks scored above 60.

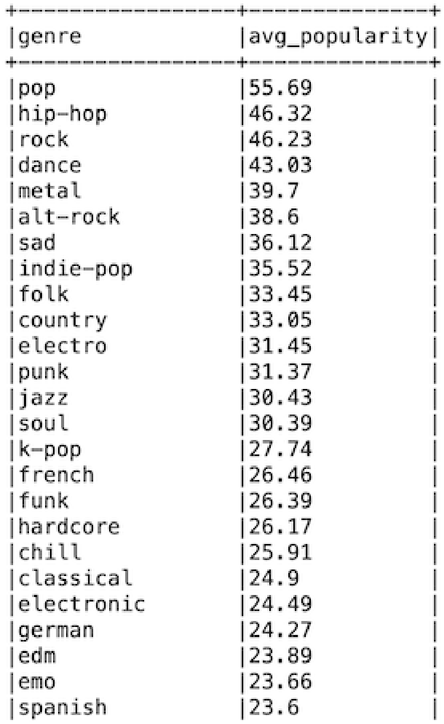


  - Over 40k tracks had a popularity score between 0-10. Like Dataset 1, this reflects a significant portion of the data. In this case, it reflects about 1/4 of it. Inversely, just over 3k tracks score above 60.


Average Song Duration from 2000-2023
Average Duration in Minutes
4.4
4.2
4.0
3.8
3.6
2000
2005
2010
2015
2020
Year


Average Song Duration from 1921-1999
Average Duration in Minutes
4.50
4.25
4.00
3.75
3.50
3.25
3.00
2.75
1920
1930
1940
1950
1960
1970
1980
1990
2000
Year

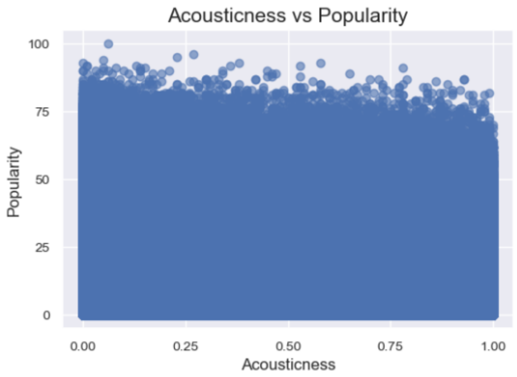
The Top 25 Genres from 2000-2023:

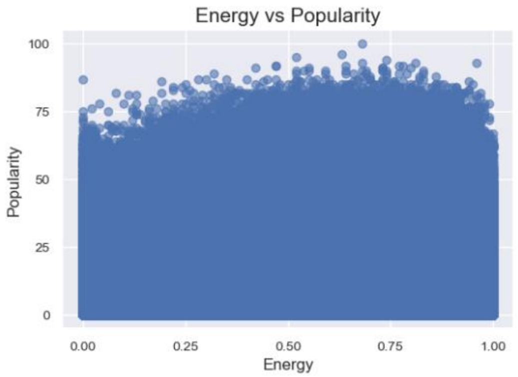


```
+------------------+--------------+
|genre             |avg_popularity|
+------------------+--------------+
|pop               |55.69         |
|hip-hop           |46.32         |
|rock              |46.23         |
|dance             |43.03         |
|metal             |39.7          |
|alt-rock          |38.6          |
|sad               |36.12         |
|indie-pop         |35.52         |
|folk              |33.45         |
|country           |33.05         |
|electro           |31.45         |
|punk              |31.37         |
|jazz              |30.43         |
|soul              |30.39         |
|k-pop             |27.74         |
|french            |26.46         |
|funk              |26.39         |
|hardcore          |26.17         |
|chill             |25.91         |
|classical         |24.9          |
|electronic        |24.49         |
|german            |24.27         |
|edm               |23.89         |
|emo               |23.66         |
|spanish           |23.6          |
```

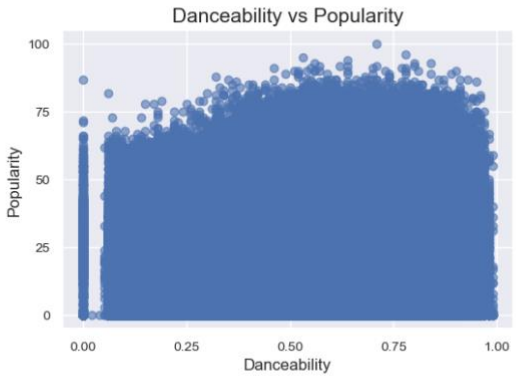
Key Feature Findings are Below:

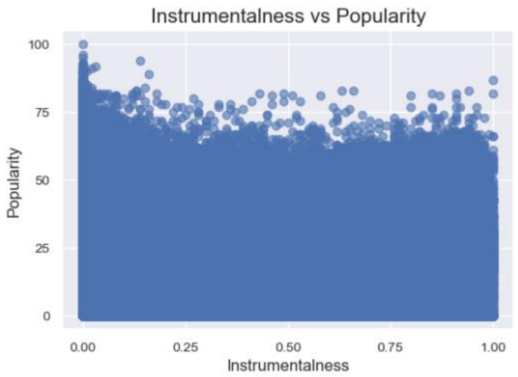
- For the 2000-2023 dataset (Dataset 1):
    - While focusing on track's/genre's respective popularities, key insights were found. The following analysis is from an "on average" perspective. The less acoustic, the more popular a track/genre is. Though, acousticness's impact on popularity is not strong. Energy's popularity sweet spot, which is measured between 0-1, hovers around 0.50-0.75. The more danceable a track/genre, the more popular... 0.55-0.75 is the sweet spot. Valence is not a strong predictor of popularity while on the other hand, speechiness is. The less speechiness, the more popular (sweet spot is 0-0.10). Loudness also is a strong predictor of popularity as its sweet spot is -15 to -5 dBs. Lastly, instrumentalness is like valence where it's not a strong predictor of popularity, though, an instrumentalness of 0 is most popular.


Acousticness vs Popularity
Popularity
100
75
50
25
0
0.00
0.25
0.50
0.75
1.00
Acousticness

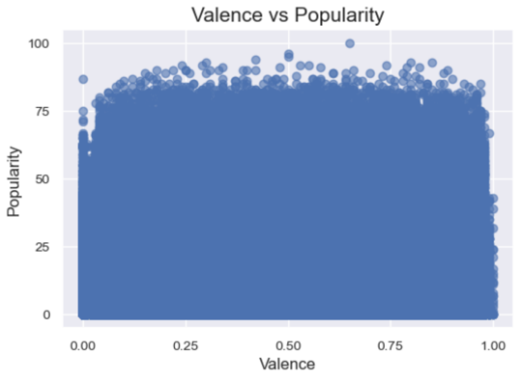

Energy vs Popularity
Popularity
100
75
50
25
0
0.00
0.25
0.50
0.75
1.00
Energy

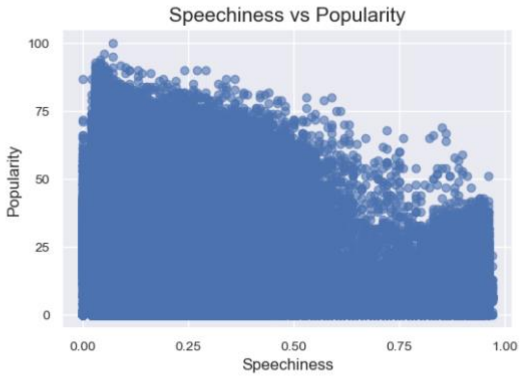
○


Danceability vs Popularity
Popularity
100
75
50
25
0
0.00
0.25
0.50
0.75
1.00
Danceability

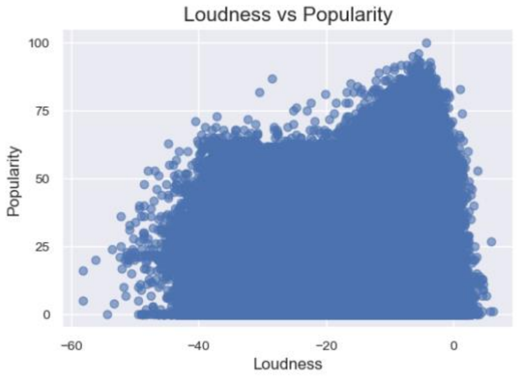

Instrumentalness vs Popularity
Popularity
100
75
50
25
0
0.00
0.25
0.50
0.75
1.00
Instrumentalness

○

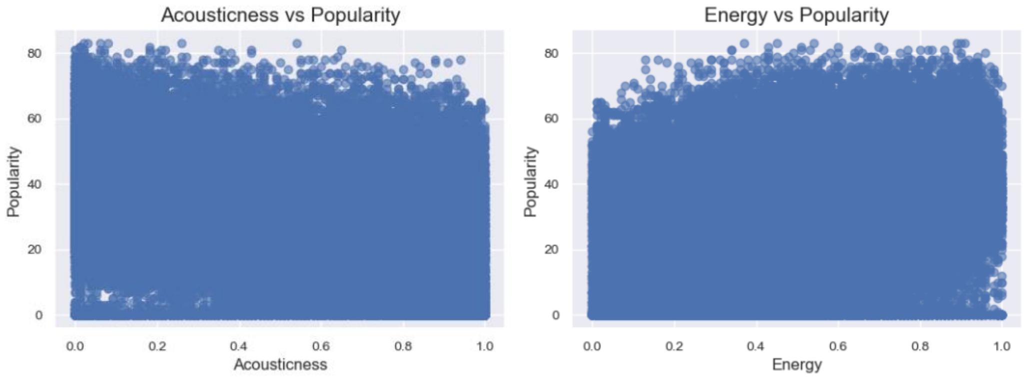

Valence vs Popularity
Popularity
100
75
50
25
0
0.00
0.25
0.50
0.75
1.00
Valence

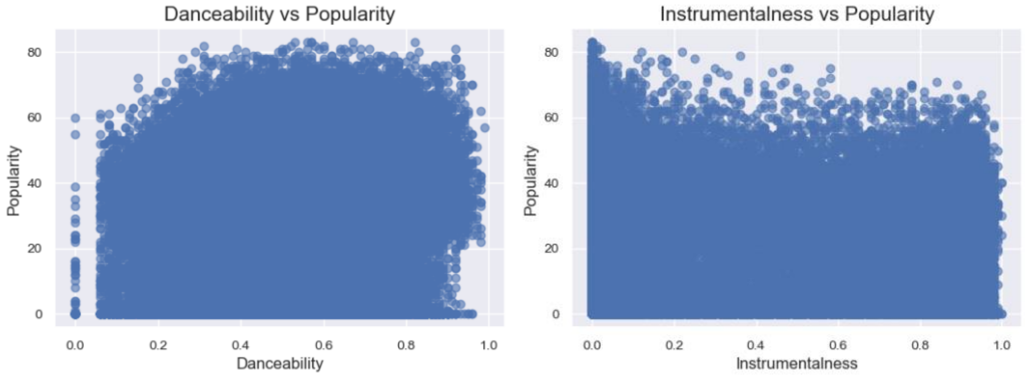

Speechiness vs Popularity
Popularity
100
75
50
25
0
0.00
0.25
0.50
0.75
1.00
Speechiness

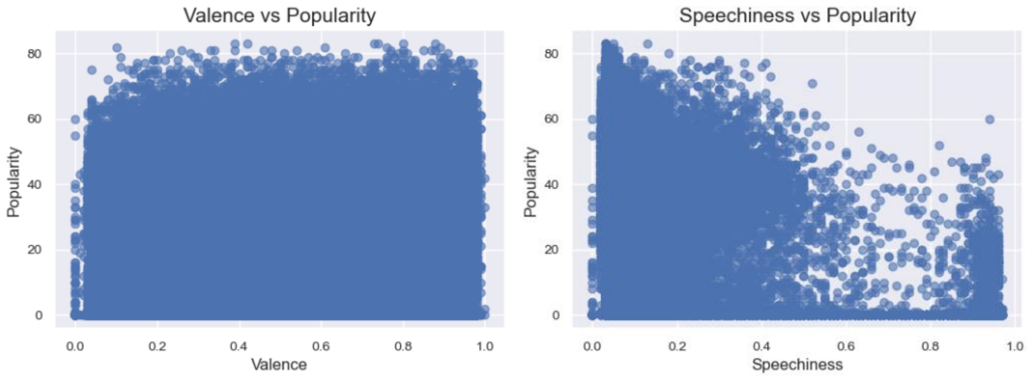
○

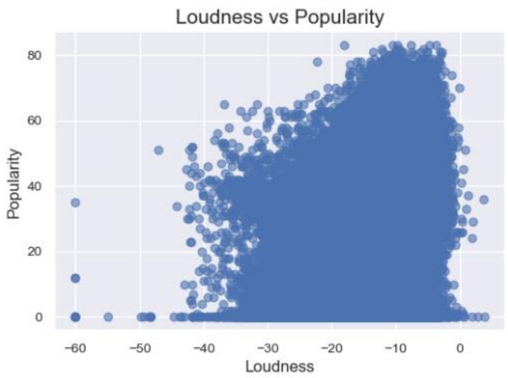

Loudness vs Popularity
Popularity
100
75
50
25
0
−60
−40
−20
0
Loudness

○

- For the 1921-1999 dataset (Dataset 2):
  - While focusing on track's/genre's respective popularities, key insights were found. The following analysis is from an "on average" perspective. The less acoustic, the more popular a track/genre is. Though, acousticness's impact on popularity is not strong. Energy's popularity sweet spot, which is measured between 0-1, hovers around 0.65-0.95 which is greater than the file_2000_2023_spotify dataframe. The more danceable a track/genre, the more popular... 0.55-0.75 is the sweet spot just like the file_2000_2023_spotify dataframe. Valence is not a strong predictor of popularity while on the other hand, speechiness is. The less speechiness, the more popular (sweet spot is 0-0.05). Loudness also is a strong predictor of popularity as its sweet spot is -15 to -5 dBs. Lastly, instrumentalness is like valence where it's not a strong predictor of popularity, though, an instrumentalness of 0 is most popular.
  - 

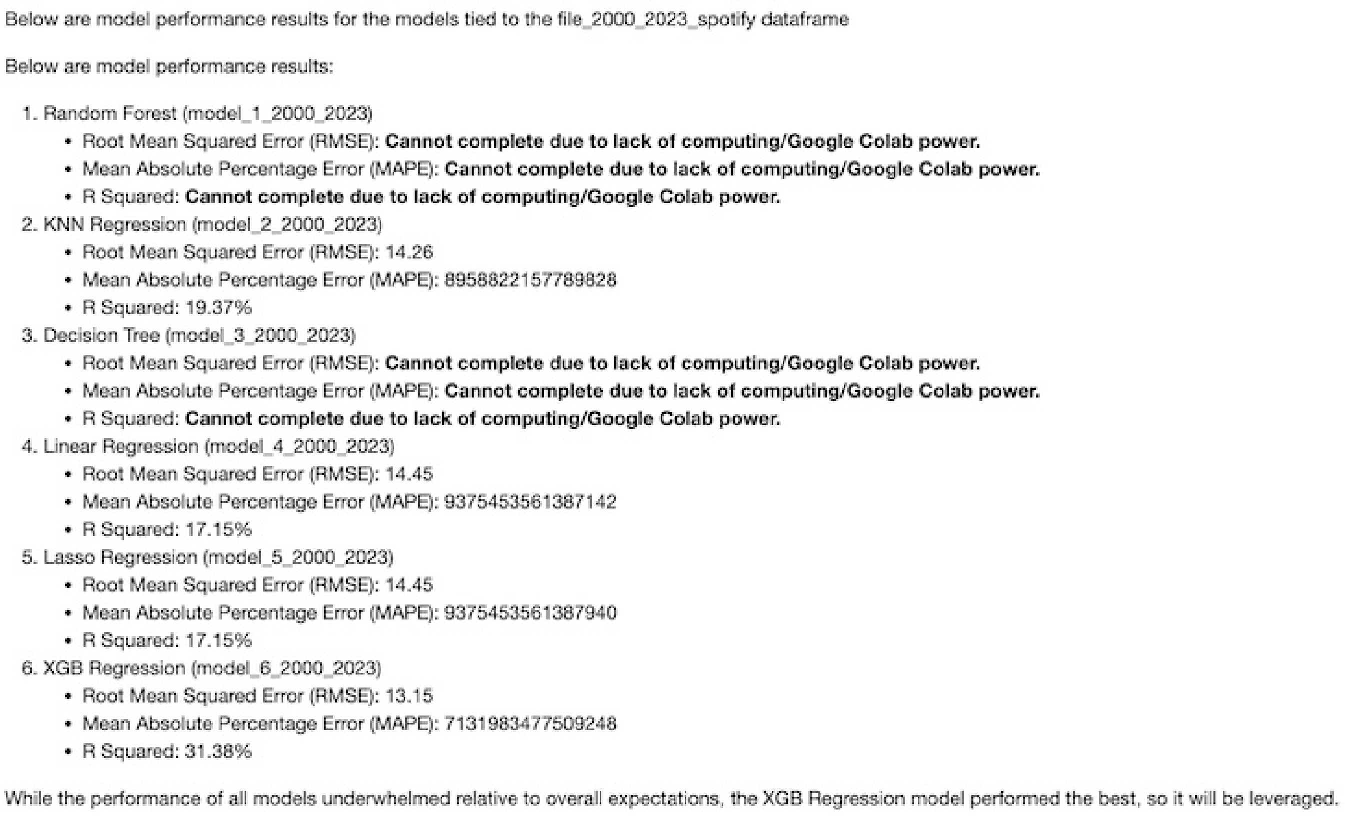
  - 

- 

- 

Modeling

The dataset lends itself to supervised machine learning and specifically, regression analysis. With that in mind, I focused on Random Forest Regression, KNN Regression, Decision Tree Regression, Linear Regression, Lasso Regression, and XGB Rgression. For Dataset 1, I dropped features 'artist_name', 'genre', 'time_signature', 'track_id', 'track_name' and for Dataset 2, I dropped features 'artist_name', 'explicit', 'id', 'track_name.' I tested the datasets on six models by allocating a 20%/80% test, train split with my target variable being the feature 'popularity.' Below are the model results:

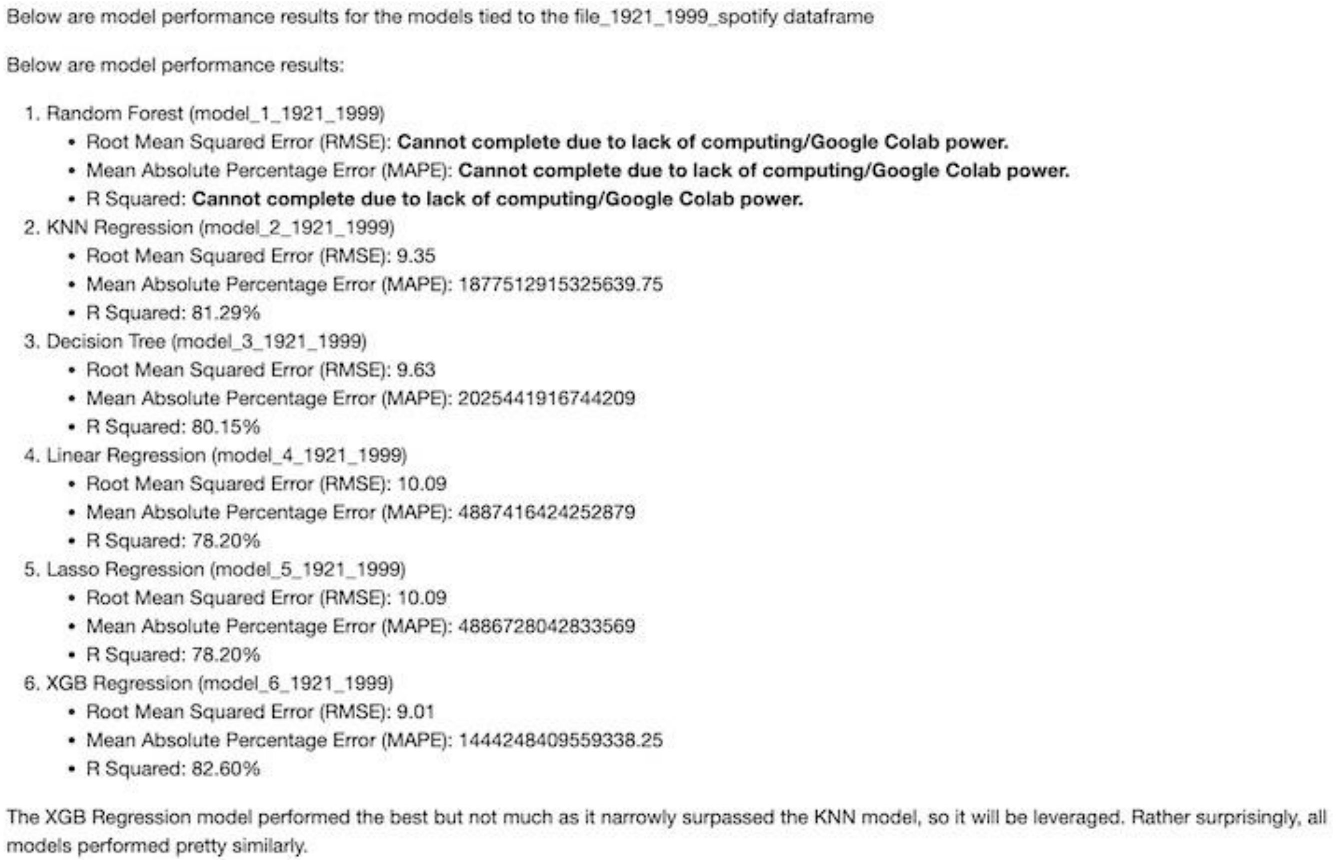


Below are model performance results for the models tied to the file_2000_2023_spotify dataframe

Below are model performance results:

1. Random Forest (model_1_2000_2023)
   - Root Mean Squared Error (RMSE): **Cannot complete due to lack of computing/Google Colab power.**
   - Mean Absolute Percentage Error (MAPE): **Cannot complete due to lack of computing/Google Colab power.**
   - R Squared: **Cannot complete due to lack of computing/Google Colab power.**
2. KNN Regression (model_2_2000_2023)
   - Root Mean Squared Error (RMSE): 14.26
   - Mean Absolute Percentage Error (MAPE): 8958822157789828
   - R Squared: 19.37%
3. Decision Tree (model_3_2000_2023)
   - Root Mean Squared Error (RMSE): **Cannot complete due to lack of computing/Google Colab power.**
   - Mean Absolute Percentage Error (MAPE): **Cannot complete due to lack of computing/Google Colab power.**
   - R Squared: **Cannot complete due to lack of computing/Google Colab power.**
4. Linear Regression (model_4_2000_2023)
   - Root Mean Squared Error (RMSE): 14.45
   - Mean Absolute Percentage Error (MAPE): 9375453561387142
   - R Squared: 17.15%
5. Lasso Regression (model_5_2000_2023)
   - Root Mean Squared Error (RMSE): 14.45
   - Mean Absolute Percentage Error (MAPE): 9375453561387940
   - R Squared: 17.15%
6. XGB Regression (model_6_2000_2023)
   - Root Mean Squared Error (RMSE): 13.15
   - Mean Absolute Percentage Error (MAPE): 7131983477509248
   - R Squared: 31.38%

While the performance of all models underwhelmed relative to overall expectations, the XGB Regression model performed the best, so it will be leveraged.



Below are model performance results for the models tied to the file_1921_1999_spotify dataframe

Below are model performance results:

1. Random Forest (model_1_1921_1999)
   - Root Mean Squared Error (RMSE): **Cannot complete due to lack of computing/Google Colab power.**
   - Mean Absolute Percentage Error (MAPE): **Cannot complete due to lack of computing/Google Colab power.**
   - R Squared: **Cannot complete due to lack of computing/Google Colab power.**
2. KNN Regression (model_2_1921_1999)
   - Root Mean Squared Error (RMSE): 9.35
   - Mean Absolute Percentage Error (MAPE): 1877512915325639.75
   - R Squared: 81.29%
3. Decision Tree (model_3_1921_1999)
   - Root Mean Squared Error (RMSE): 9.63
   - Mean Absolute Percentage Error (MAPE): 2025441916744209
   - R Squared: 80.15%
4. Linear Regression (model_4_1921_1999)
   - Root Mean Squared Error (RMSE): 10.09
   - Mean Absolute Percentage Error (MAPE): 4887416424252879
   - R Squared: 78.20%
5. Lasso Regression (model_5_1921_1999)
   - Root Mean Squared Error (RMSE): 10.09
   - Mean Absolute Percentage Error (MAPE): 4886728042833569
   - R Squared: 78.20%
6. XGB Regression (model_6_1921_1999)
   - Root Mean Squared Error (RMSE): 9.01
   - Mean Absolute Percentage Error (MAPE): 1444248409559338.25
   - R Squared: 82.60%

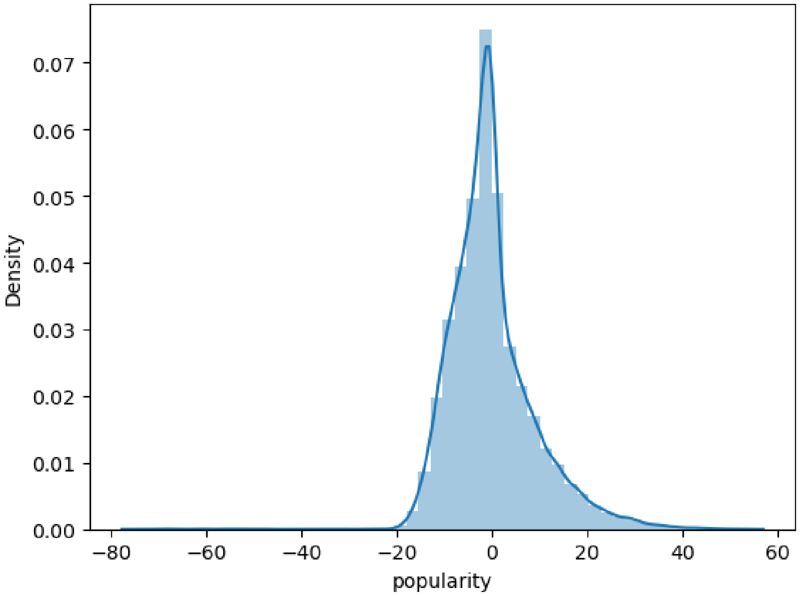
The XGB Regression model performed the best but not much as it narrowly surpassed the KNN model, so it will be leveraged. Rather surprisingly, all models performed pretty similarly.

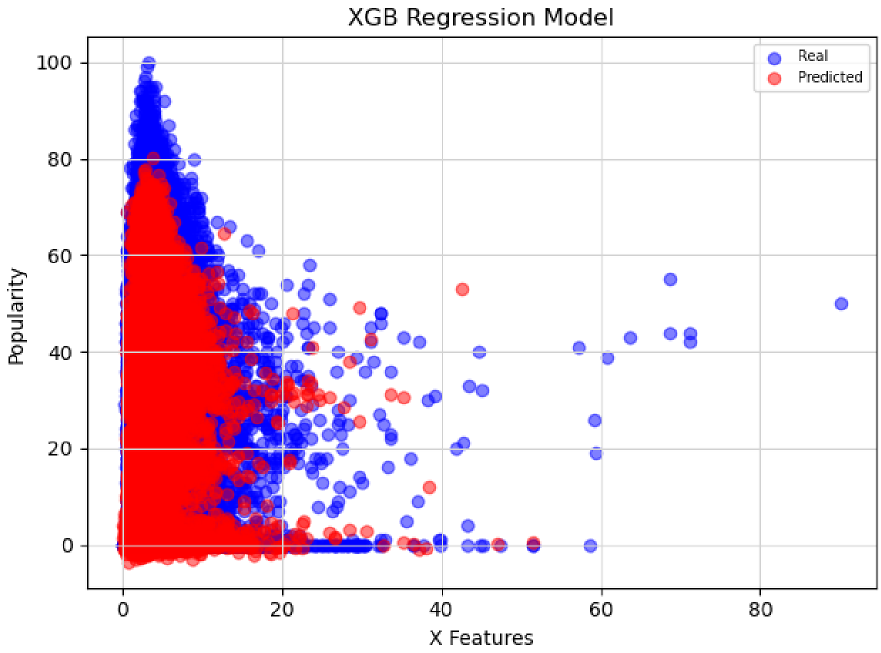
I focused on RMSE, MAPE, and R Squared as my metrics to give me a more thorough understanding of the quality of the respective model performances. RMSE determines the absolute fit of the model to the data as it measures the average difference between a model's predicted values and its actual values. MAPE considers the sum of the individual absolute errors divided by the demand (each period separately). It is the average of the percentage errors. For both RMSE and MAPE, the lower the value the better. Finally, R Squared measures the proportion of variance in the dependent variable that can be explained by the independent variable. The greater the value the better. Vaguely, 0.65/65% or greater is considered pretty good.

Modeling

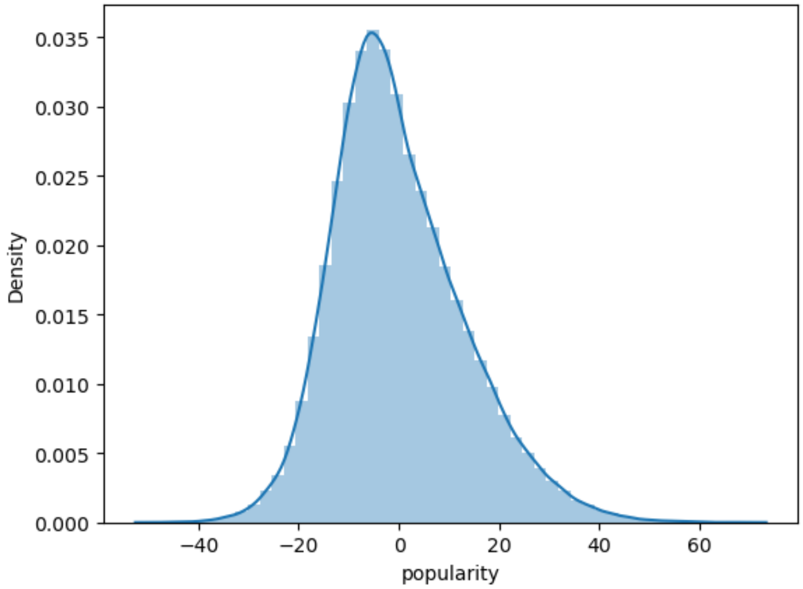
For each dataset, the XGB models performed best, so they were leveraged for predictive analysis. Their respective results are below. The first graph for each model shows the difference between the actual dependent feature vs the predicted dependent feature.

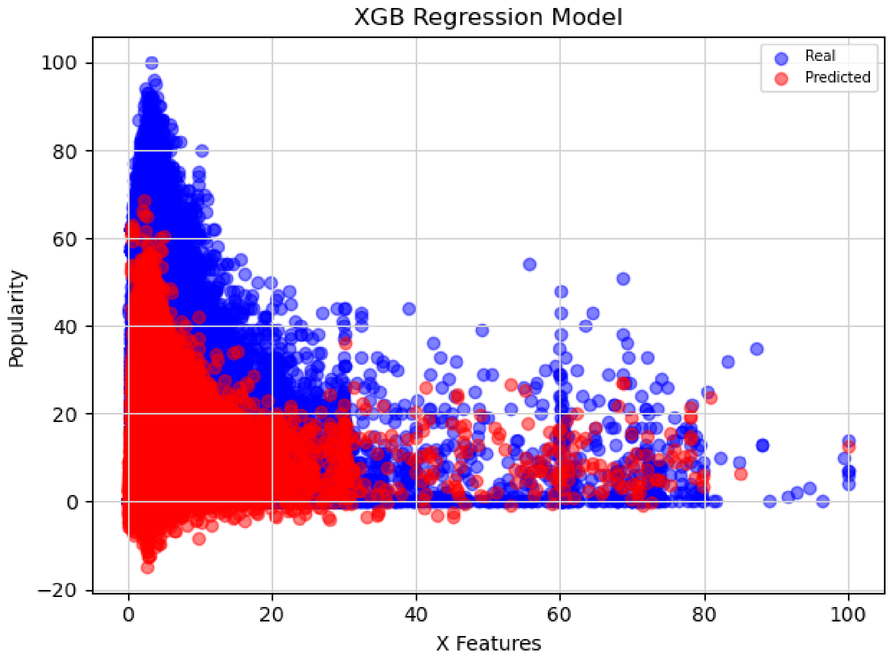
The model tied to Dataset 1 performs well at predicting 'popularity' by using all features tied to the X variable.





The model tied to Dataset 2 does not perform well at predicting 'popularity' by using all features tied to the X variable.





Recommendations

<u>Given the music industry's fluctuation over time:</u>

- Stay up to date on trends and shifts in the industry year-over-year... it moves faster these days.
  - Remain flexible so one can cater more efficiently to trends and shifts.
- Cater venues/shows to feature the most popular artists to ensure profitability benchmarks are hit... so provide your target audience(s) with who they want, if possible.
  - Explore merchandise for most popular artists as well.
- Understand that track/artist/genre popularity varies by generation and that popularity is affected by numerous features varying from danceability, energy, loudness, valence and so on.

Future Work

- Hit Spotify's API to pull more data tied to the years 1921-1999 (Dataset 2). The primary goal would be to have scale like Dataset 1 where the row count is over one million and where each line item is grouped by genre like Dataset 1. This thought assumes the results would be better than Dataset 2 as is.
- Explore impact of cultural changes over time to understand its impact on genre popularity.

- Leverage an additional dataset to explore which regions or states in the U.S. or broader scale... which countries prefer which genres over time.
- Leverage an additional dataset, if available, to explore social media's impact on genre or artist popularity over time.
- For artists seeking to craft a popular track, be mindful of the most popular recent genres which are pop, hip-hop, dance and house among a few others. Being mindful of strong feature correlations such as energy and loudness (strong positive) and acousticness and energy or loudness (strong negative) will also aid in your creation process.