Spotify Streaming Analysis Report

Introduction:

This dataset exhibits a collection of metadata of Spotify tracks of various music genres. Metadata such as artist(s) name, album, track title, popularity score, duration, and whether a track is marked as explicit. Additionally, it provides audio attributes such as danceability, energy, acousticness, instrumentalness, and tempo therefore characterizing songs beyond surface-level descriptions, which allows for a deeper and more insightful understanding and analysis.

The main goal of this project is to analyze patterns in a song's production, popularity, and traits that would apply to certain genres. Some questions that this analysis attempts to answer are:

- How do various audio features correlate with a track's popularity on Spotify?
- What audio characteristics are most representative of different musical genres?
- What makes a popular song *popular*?

After addressing these questions, this analysis offers insights to the evolution of music streaming. Following this report, it will inform those that are passionate about music about trends, preferences, and other musical statistics that are to be found.

Exploratory Analysis:

Initial Data Overview:

- The initial dataset comprised 114,000 entries and 21 columns.
- The dataset included one Boolean, nine Float, six Integer, and five Object type columns.
- One missing value was identified within the dataset.

Handling Missing Values and Column Information:

- A single row, corresponding to Token 65900, was found to have incomplete fundamental song information, specifically the artist's name, album name, and track name. Consequently, this incomplete row was removed from the dataset.
- Following the removal of the row with the missing value, the dataset now contains 113,999 entries.
- The 'Unnamed: 0' column was determined to be non informative and was therefore excluded from the analysis.
- The dataset initially appeared to contain duplicate entries, which typically necessitate removal to avoid redundancy.
- However, further investigation revealed that these apparent duplicates originated from the 'track_genre' column, where individual songs could be associated with multiple genres. Therefore, these entries were retained to preserve the multi genre representation of the tracks.

Analysis of Numeric Variables and Visualization:

- A correlation analysis was performed on the continuous variables:
  - *danceability*
  - *energy*
  - *acousticness*
  - *tempo*
  - *loudness*
  - *speechiness*
  - *valence*

Insights from the Continuous Variable Heatmap:

- A positive correlation was observed between *valence* (indicating positivity) and *danceability*, suggesting that more positive songs tend to be more danceable.
- *Energy* and *loudness* exhibited a positive relationship, where louder songs generally possess higher energy levels.
- Conversely, *acousticness* showed a negative correlation with *energy*, implying that more acoustic songs tend to have lower energy.
- For the most part, *acousticness* and *loudness* were negatively correlated, suggesting that acoustic tracks are generally quieter.
- *Speechiness*, which measures the presence of spoken words, displayed a weak positive correlation with both *danceability* and *energy*, indicating a slight tendency for songs with spoken elements to be more danceable or energetic, although this relationship is not strong.

Relationship between Average Loudness and Energy Levels:

- A line plot indicates a generally logarithmic increase in average loudness as energy levels rise.

Relationship between Average Acousticness and Energy Levels:

- The line plot reveals an inverse relationship, with average acousticness decreasing as energy levels increase.

Relationship between Average Loudness and Acousticness Levels:

- The line plot shows a general decline in loudness with higher acousticness, punctuated by a significant drop in decibels at the higher end of the acousticness spectrum.

Relationship between Average Danceability and Valence Levels:

- The line plot illustrates an upward trend, demonstrating that average danceability increases with higher valence scores.

Relationship between Average Loudness and Tempo Levels:

- The line plot indicates that average loudness generally increases with tempo, exhibits a slight dip, and then concludes with a significant decrease at the highest tempos.

Distribution of Numeric Variables:

- The distribution of danceability appeared relatively uniform, which is expected given the diverse nature of music, encompassing tracks intended for dancing and those for listening.
- The energy distribution showed a general upward trend, suggesting a prevalence of songs with progressively higher energy levels within the dataset.
- The dataset contained a limited number of acoustic songs, as indicated by the distribution.
- The tempo of the songs was widely distributed, reflecting a variety of paces across the dataset.
- The loudness distribution exhibited a bell shaped curve, likely influenced by the characteristics of different music genres.
- The valence distribution was generally high across most counts, indicating a prevalence of songs with positive sentiment.
- The presence of outliers in the visualizations and distributions did not appear to significantly skew the overall data trends and thus did not fundamentally alter the conclusions drawn.

Analysis of Categorical Values:

- An analysis of the top and bottom 10 occurrences within the *'track_genre'*, *'artists'*, and *'album_name'* columns was conducted.

  Most Frequent Occurrences:

  - For *track_genre*, given the large dataset size (over 100,000 entries), the top 10 most frequent genres, as depicted in the bar chart, all had value counts exceeding 1000. These genres, presented alphabetically, range from "acoustic" to "brazil".
  - For *artists*, the high frequency of certain artists was notably linked to the number of genres associated with their tracks, as evidenced by the difference between "Unique Tracks" and "Total Tracks":

Artist	**Career Span**	**Unique Tracks**	**Total Tracks**
The Beatles	1960–1970	149	279
George Jones	1953–2013	215	271
Stevie Wonder	1961–present	16	236
Linkin Park	1996–2017, 2023–present	84	224
Ella Fitzgerald	1934–1993	18	222
Prateek Kuhad	2011–present	46	217
Feid	2013–present	12	202
Chuck Berry	1953–2017	13	190
Håkan Hellström	1991–present	122	183
OneRepublic	2002–present	59	181

Table 1: Most Frequent Artists: Career Spans, Unique Tracks, and Total Tracks

Initially, it was hypothesized that artists with longer career spans would naturally accumulate more tracks. However, further investigation revealed that the total track count for an artist in this dataset is more strongly correlated with the number of genres associated with their songs, as previously mentioned.

  - Interestingly, the most frequently occurring *album_name* entries appeared to be predominantly seasonal albums rather than general releases.

- Visualizing the least frequent counts for *'track_genre'*, *'artists'*, and *'album_name'* proved unhelpful due to the large number of unique entries, with the majority having a frequency of only 1. This longtail distribution made a meaningful "bottom 10" visualization impractical.

Higher Level Analysis:

1. Top 10 Genres by Average Popularity:

A key question to address was identifying the top 10 most "popular" genres within this dataset. The exploratory analysis revealed a substantial number of data entries for various genres. For artists, the number of *unique tracks* for the most frequent artists was comparatively lower than their total track count due to multi genre classifications.

Utilizing a pie chart to illustrate the top 10 genres by average popularity as percentages, it is evident that `pop-film` is the most popular genre, followed by `k-pop`, and so forth.

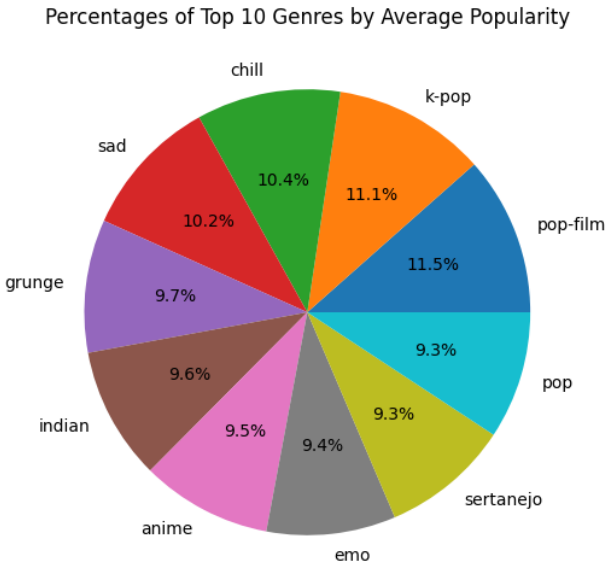

Figure 1: Top 10 Genres by Average Popularity

2. Top 10 Artists by Average Popularity:

Following the analysis of genres, identifying the artists with the highest average popularity was another important aspect to consider.

Again, employing a pie chart to visualize percentage values, the following artists emerged as the top 10 in terms of average popularity.

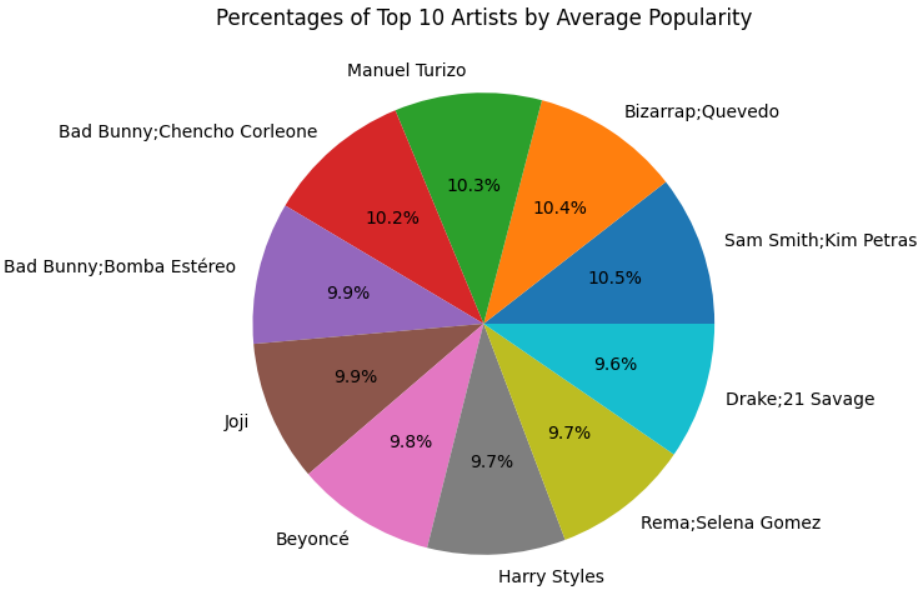

Figure 2: Top 10 Artists by Average Popularity

Artist(s)	**Genre(s)**
Sam Smith;Kim Petras	Dance & Pop
Bizarrap; Quevedo	Hip-Hop
Manuel Turizo	Latin, Latino, reggae, reggaeton
Bad Bunny; Chencho Corleone	Latin, Latino, reggae, reggaeton
Bad Bunny; Bomba Estéreo	Latin, Latino, reggae, reggaeton
Joji	Pop
Beyoncé	Dance
Harry Styles	Pop
Rema; Selena Gomez	Pop
Drake; 21 Savage	Hip-Hop

Table 2: Top 10 Artists Genres

3. Linear Regression: Speechiness vs Tempo:

Null Hypothesis
These two categories are uncorrelated.

Alternative Hypothesis
The correlation among the variables are nonzero.

The slope of between the variables is a positive trend of **4**, (r=0.017, $p < 0.01$). However as seen in the heatmap before, this is a strongly weak positive correlation. But the p-value is statistically significant as it can have impact on a song. Discovering this, the null hypothesis can be rejected.

4. K-Means Clustering:

cluster	0	1	2	3	4	5	6	7	8
danceability	0.530549	0.670655	0.347784	0.573955	0.472004	0.470193	0.586176	0.522175	0.716733
energy	0.738668	0.507735	0.169670	0.670453	0.815388	0.313206	0.744782	0.758014	0.768906
loudness	-6.119909	-9.221327	-21.373121	-11.163560	-5.236839	-11.710334	-8.409667	-6.984827	-5.797511
speechiness	0.067194	0.081390	0.050886	0.836033	0.103635	0.044713	0.070921	0.085794	0.092800
acousticness	0.102983	0.588553	0.863385	0.732562	0.111268	0.747609	0.104530	0.281898	0.137521
instrumentalness	0.028417	0.034802	0.827048	0.009980	0.036198	0.031206	0.797010	0.068894	0.021262
liveness	0.177058	0.164539	0.161280	0.652586	0.196743	0.159251	0.170111	0.753127	0.167646
valence	0.316666	0.657978	0.185566	0.442053	0.477254	0.298959	0.332750	0.509049	0.725386
tempo	109.941132	117.154248	103.203736	101.238308	164.595194	112.773173	126.815542	123.602399	116.373098
explicit	0.109307	0.075472	0.002445	0.560550	0.110468	0.021264	0.036178	0.057371	0.126732

Figure 3: K-Means Cluster Results

Cluster	Interpreted Genre
Cluster 0	Electronic Dance Music
Cluster 1	Chill or Mellow
Cluster 2	Instrumental Acoustic
Cluster 3	Spoken Word or Hip-Hop
Cluster 4	High-Energy Pop, Electronic Dance, or Rock-Pop
Cluster 5	Soft Acoustic
Cluster 6	Electronic Instrumental
Cluster 7	Live Performance/Energetic Recording
Cluster 8	Club Hits/Feel-Good Dance

Table 3: Interpreted Music Genres by Cluster

`Cluster 0:`

- Moderate Danceability (0.53)
- High Energy (0.73)
- Decent Valence (0.31)
- Low Acousticness (0.10)
- Very Low Instrumentalness (0.02)
- **Interpretation:** Likely Electronic Dance Music due to its high energy, moderate danceability, and low acoustic/instrumental presence.

`Cluster 1:`

- Moderate Danceability (0.67)
- Moderate Energy (0.50)
- Low Loudness (-9.22)
- Moderate Acousticness (0.58)
- **Interpretation:** Suggests Chill or Mellow music characterized by moderate danceability and acousticness, and lower energy/loudness.

`Cluster 2:`

- High Acousticness (0.86)
- High Instrumentalness (0.82)
- Very Low Loudness (-21)
- Low Energy (0.16)
- **Interpretation:** Points to Instrumental Acoustic music, evident from the high acousticness and instrumentalness coupled with low energy and loudness.

`Cluster 3:`

- Moderate Danceability (0.57)
- Moderate-High Energy (0.67)
- Medium Loudness (-11.16)
- High Speechiness (0.84)
- **Interpretation:** Indicates Spoken Word or Hip-Hop, defined by high speechiness and moderate danceability/energy.

`Cluster 4:`

- High Energy (0.82)
- Fairly High Loudness (-5.24)
- Moderate Danceability (0.47)
- Low Acousticness (0.11)
- Low Instrumentalness (0.03)
- **Interpretation:** Could be High-Energy Pop, Electronic Dance, or Rock-Pop, given the high energy/loudness and low acoustic/instrumental qualities.

`Cluster 5:`

- Low Danceability (0.47)
- Low Energy (0.31)
- Medium Loudness (-11.71)
- High Acousticness (0.75)
- **Interpretation:** Likely Soft Acoustic music, characterized by low danceability and energy alongside high acousticness.

`Cluster 6:`

- High Energy (0.74)
- Moderate Danceability (0.59)
- Mid-Range Loudness (-8.41)
- High Instrumentalness (0.80)

- **Interpretation:** Suggests Electronic Instrumental music due to its high energy and instrumentalness with moderate danceability.

`Cluster 7:`

- Moderate Danceability (0.52)
- High Energy (0.76)
- Mid-Range Loudness (-6.98)
- High Liveness (0.75)
- **Interpretation:** Indicates Live Performances or energetic recording sessions, marked by high energy and liveness with moderate danceability.

`Cluster 8:`

- Highest Danceability (0.72)
- Highest Energy (0.77)
- High Loudness (-5.80)
- Low Acousticness (0.14)
- High Valence (0.73)
- **Interpretation:** Points to Club Hits or feel-good dance tracks, characterized by high danceability, energy, loudness, and valence, with low acousticness.

5. Explicit vs Non-Explicit: Popularity Comparison:

By conducting a T-Test, seeing the difference between explicit and non-explicit popular songs will show if explicit words matter.

Null Hypothesis:
The means of these two groups are the same with each other.

Alternative Hypothesis:
The means of these two groups are different.

First evaluating if there are more popular explicit songs on average with non-explicit was conducted. For explicit, a mean of 36.45 was produced and for non-explicit a mean of 32.93 was produced. Looking at the averages at face value, they difference is relatively close, to verify this assertion the T-Test was conducted.

Following the T-test, the statistic value is 14.89, suggesting a big difference. A P-Value less than 0.01 ($3.85500797421256 \times 10^{-50}$). After this, the null hypothesis can be rejected as explicit songs tend to be more popular than non-explicit songs.

6. Linear Regression: Danceability vs Tempo:

Null Hypothesis
These two categories are not correlated.

Alternative Hypothesis
The correlation among the variables are nonzero.

The slope of between the variables is a negative trend of $-\mathbf{8}$, ($r=-0.05$, $p < 0.01$). However as seen in the heatmap before, this is a weak negative correlation to almost no correlation. But the p-value is statistically significant as it can have impact on how someone can dance along to the tempo of a song. The null hypothesis can be rejected.

Conclusion:

After thorough investigation, several findings have been uncovered.

Firstly, while comparing pie charts for "Top 10 genres among popularity" and "Top 10 artists among popularity", different genres were listed in each chart, which can be a reflection of a listener's preferences.

Moreover, both linear regression charts portray correlations more effectively than in the heatmap and how statistically significant it is.

Such as in 'Speechiness vs Tempo', the P-Value showed how significant these two variables can affect a song even with a weak correlation.

Similarly in 'Danceability vs Tempo', a weak correlation was shown in the correlation heatmap but having a significant P-Value illustrated that there is significance to these two variables.

The clustering analysis revealed distinct song genres and what "creates" that song genre through audio features and other variables.

Finally the explicit vs non-explicit relationship portrayed preferences among listeners for a song's popularity. The findings revealed explicit songs were more popular than that of non-explicit.

All in all, this dataset illustrates the ability and how profound data analysis can be not just for music trends but for any dataset. Allowing one to define variables to interpret their meanings and discover correlations between the variables created, it allowed understanding for a listener's preferences to help create recommendation systems, production, and more.