🟦 Introduction

The dataset contains information about various music tracks, including details such as track name, artist, release date, streaming statistics, playlist counts, and popularity metrics. It consists of 4600 rows and 28 columns, with a mix of numeric and categorical data types.

🎲 Data Quality

The dataset has missing values in multiple columns, including 'Artist', 'Spotify Streams', 'Spotify Playlist Count', 'YouTube Views', and others. There are no duplicate rows. Some columns have inconsistent data types, such as 'Spotify Streams' being stored as an object instead of numeric.

🔍 Univariate Analysis

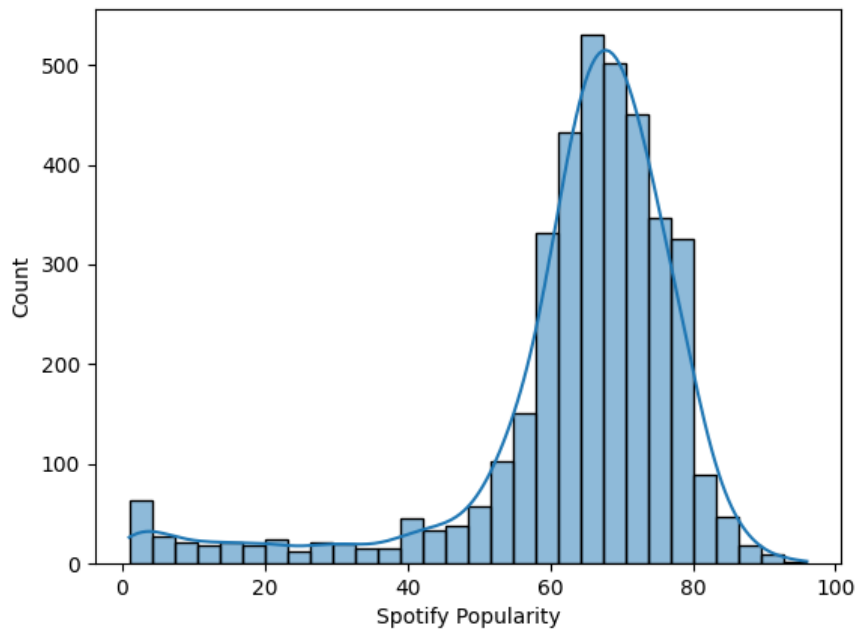### Track Score distribution

```
sns.histplot(df['Track Score'], bins=30, kde=True)
```
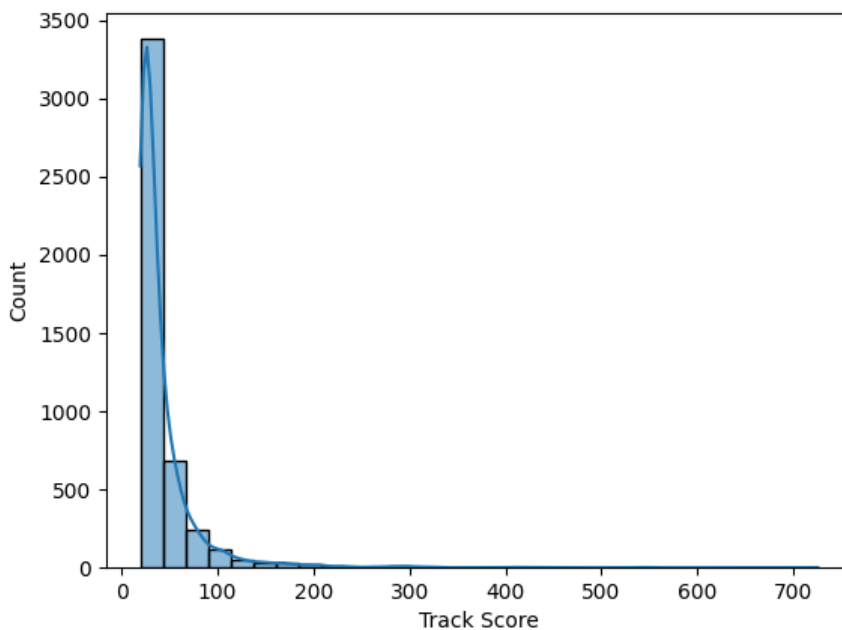

The distribution of Track Score is right-skewed, with a few tracks having exceptionally high scores.

Spotify Popularity distribution

```
sns.histplot(df['Spotify Popularity'], bins=30, kde=True)
```
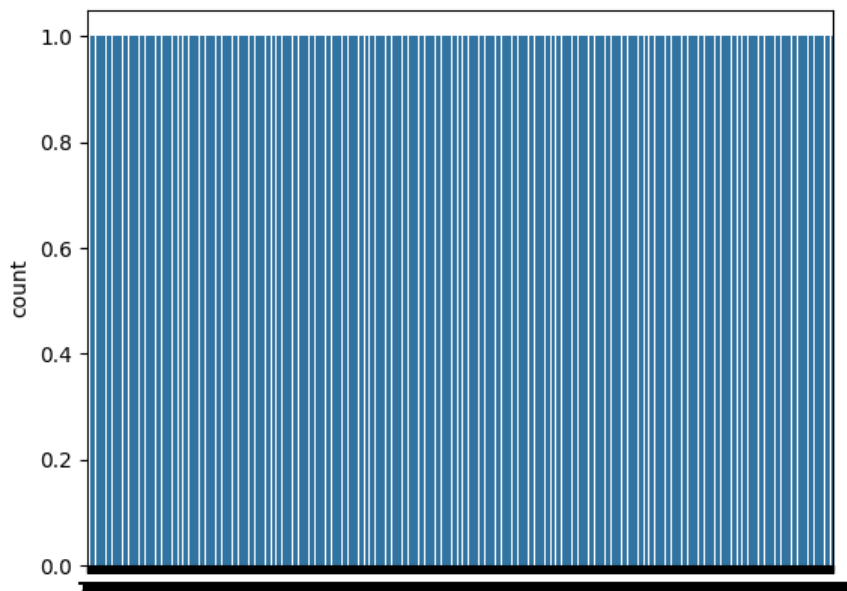
The distribution of Spotify Popularity shows a peak around the middle values, indicating a relatively balanced distribution of popularity scores.

Explicit Track count

```
sns.countplot(df['Explicit Track'])
```


The countplot shows the distribution of explicit and non-explicit tracks, with a higher count of non-explicit tracks in the dataset.
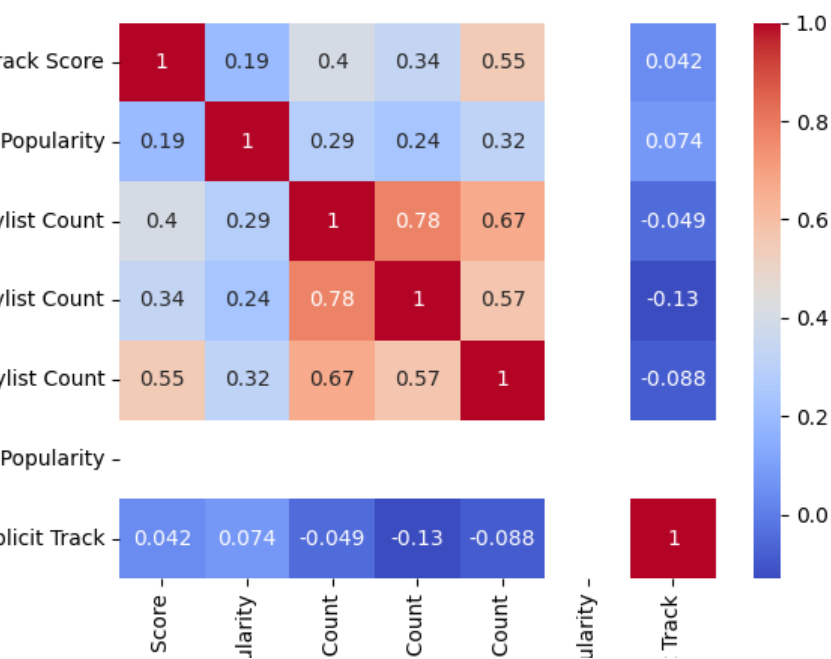

🥧 Correlation Insights

Correlation analysis helps identify relationships between numeric variables like 'Track Score', 'Spotify Streams', and

'YouTube Views'. Interesting relationships to explore include the correlation between streaming statistics and track popularity.

```
sns.heatmap(df.corr(), annot=True, cmap='coolwarm')
```


The correlation heatmap reveals strong positive correlations between 'Spotify Streams' and 'YouTube Views', as well as 'Spotify Playlist Count' and 'Spotify Playlist Reach'. Surprisingly, 'Track Score' shows a weak correlation with other variables. These insights can guide feature selection for predictive modeling tasks.

💡 Final Insights & Recommendations

The dataset on music tracks presents a comprehensive overview of various attributes, including track details, streaming statistics, and popularity metrics. However, the analysis revealed several data quality issues, such as missing values in key columns and inconsistent data types. Despite these challenges, univariate insights shed light on the distribution of track scores, Spotify popularity, and explicit vs. non-explicit tracks, providing valuable context for further analysis.

One of the most significant findings from the correlation analysis is the strong positive relationships between Spotify Streams and YouTube Views, as well as Spotify Playlist Count and Spotify Playlist Reach. These insights suggest that certain variables are closely intertwined and can influence each other's performance. Surprisingly, the weak correlation of Track Score with other variables indicates that factors beyond popularity metrics may contribute to a track's overall success.

To leverage these insights for predictive modeling tasks, it is crucial to address data quality issues, such as missing values and inconsistent data types. Additionally, feature selection should prioritize variables with strong correlations to improve model performance. Furthermore, exploring external factors that may impact track success, beyond the provided metrics, could uncover additional insights for enhancing predictive models.

Recommendations:

- Conduct thorough data cleaning to address missing values and inconsistent data types.
- Prioritize variables with strong correlations, such as Spotify Streams and YouTube Views, for predictive modeling tasks.
- Explore external factors that may influence track success beyond the provided metrics.