Most_streamed_Spotify_songs_2024

July 7, 2024

Introduction:

In the captivating landscape of the music industry, the most streamed Spotify songs of 2024 offer a unique window into the evolving tastes and preferences of listeners worldwide. This comprehensive dataset, meticulously curated, presents an opportunity to uncover insightful patterns and trends that have shaped the musical landscape in the recent past.

Through a deep dive into this data, we embark on a journey of exploration, uncovering the hidden narratives that lie beneath the surface. From the distribution of release dates to the interplay between various playlist metrics, this analysis promises to shed light on the factors that have propelled certain artists and tracks to the forefront of the streaming revolution.

By examining the trends and correlations within the data, we can gain a better understanding of the dynamic nature of the music industry, anticipating the shifts and innovations that will shape the future of music consumption. This exploration promises to captivate and inspire, as we uncover the profound insights hidden within the most streamed Spotify songs of 2024.

[10]:
```
# This Python 3 environment comes with many helpful analytics libraries␣
↪installed
# For example, here's several helpful packages to load
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns

# Input data files are available in the read-only "../input/" directory
# For example, running this (by clicking run or pressing Shift+Enter) will list␣
↪all files under the input directory

# Data Cleaning:
# Missing Values: Handle missing values appropriately, such as by imputing or␣
↪removing rows with missing data, depending on the context and the proportion␣
↪of missing values.
# Data Types: Convert any columns with incorrect data types.
# Duplicate Entries: Ensure that each row represents a unique song.
# Outliers: Handle any outliers in the data, such as by winsorizing or removing␣
↪them, depending on the nature of the data and the potential impact of the␣
↪outliers.
```

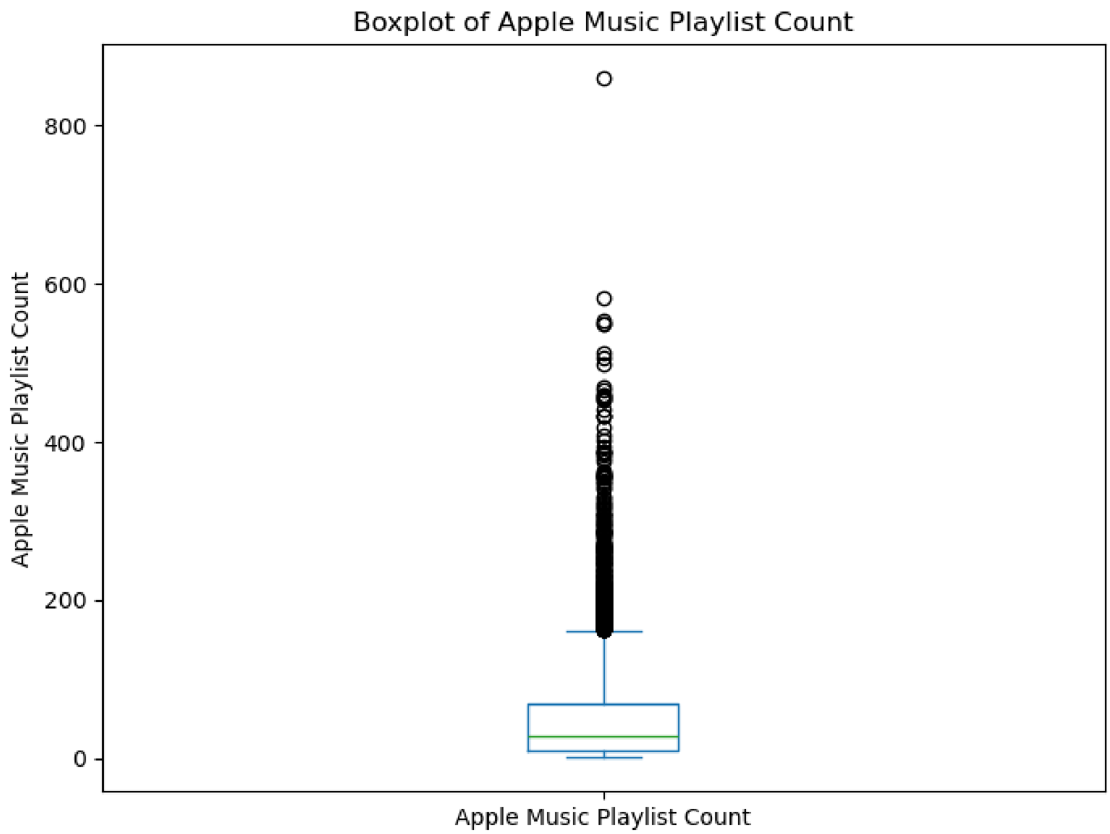
```python
# Data Formatting: Ensure data is formatted consistently, such as by standardizing date and time formats, capitalizing names, and removing any unnecessary whitespace or special characters.

# Task 1:
df = pd.read_csv(r'C:\Users\DELL\Desktop\Most Streamed Spotify Songs 2024.csv', encoding='latin-1')
missing_values = df.isnull().sum()
missing_values_df = pd.DataFrame({'Column': df.columns, 'Missing Values': missing_values})
missing_values_df = missing_values_df[missing_values_df['Missing Values'] > 0]
missing_values_df = missing_values_df.sort_values('Missing Values', ascending=False)
df = df.drop(missing_values_df.iloc[0, 0], axis=1)
df = df.col

# Task 2:
df['Spotify Playlist Count'] = df['Spotify Playlist Count'].str.replace(',', '').astype(int)

# Task 3:
df = df.drop_duplicates()

# Task 4:
track_score_iqr = df['Track Score'].quantile(0.75) - df['Track Score'].quantile(0.25)
track_score_upper = df['Track Score'].quantile(0.75) + 1.5 * track_score_iqr
track_score_lower = df['Track Score'].quantile(0.25) - 1.5 * track_score_iqr
df = df[(df['Track Score'] >= track_score_lower) & (df['Track Score'] <= track_score_upper)]
apple_music_iqr = df['Apple Music Playlist Count'].quantile(0.75) - df['Apple Music Playlist Count'].quantile(0.25)
apple_music_upper = df['Apple Music Playlist Count'].quantile(0.75) + 1.5 * apple_music_iqr
apple_music_lower = df['Apple Music Playlist Count'].quantile(0.25) - 1.5 * apple_music_iqr
df = df[(df['Apple Music Playlist Count'] >= apple_music_lower) & (df['Apple Music Playlist Count'] <= apple_music_upper)]

# Task 5:
df['Date'] = pd.to_datetime(df['Date'])
df['Month'] = df['Date'].dt.month
df['Year'] = df['Date'].dt.year

# Univariate Analysis
```

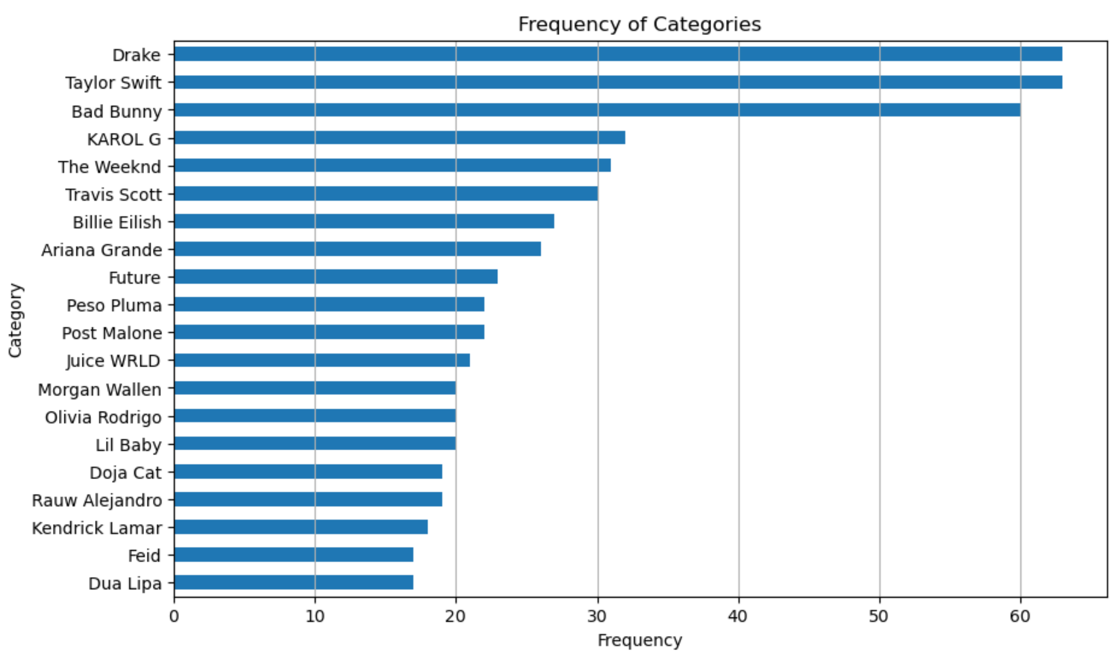
```python
# Task 6 (Numerical Univariate Analysis):
num_cols = df.select_dtypes(include='number').columns
print(df[num_cols].describe())
print(f'IQR for Track Score: {track_score_iqr}')

# Task 7 (Categorical Univariate Analysis):
cat_cols = df.select_dtypes(exclude='number').columns
for col in cat_cols:
    print(f'Value counts for {col}:')
    print(df[col].value_counts())
    plt.figure(figsize=(12, 6))
    df[col].value_counts().head(20).plot(kind='barh')
    plt.title(f'Top 20 {col}s')
    plt.show()

# Task 8 (Heatmap)
plt.figure(figsize=(12, 10))
sns.heatmap(df[num_cols].corr(), annot=True, cmap='YlOrRd')
plt.title('Correlation Heatmap')
plt.show()

# Task 9 (Scatter Plot)
plt.figure(figsize=(8, 6))
plt.scatter(df['Numerical Feature 1'], df['Numerical Feature 2'])
plt.xlabel('Numerical Feature 1')
plt.ylabel('Numerical Feature 2')
plt.title('Scatter Plot of Highly Correlated Numerical Features')
plt.show()

# Task 10 (Crosstable)
pd.crosstab(df['Categorical Feature 1'], df['Categorical Feature 2'])'
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 4600 entries, 0 to 4599
Data columns (total 29 columns):
 #   Column                      Non-Null Count  Dtype
---  ------                      --------------  -----
 0   Track                       4600 non-null   object
 1   Album Name                  4600 non-null   object
 2   Artist                      4595 non-null   object
 3   Release Date                4600 non-null   object
 4   ISRC                        4600 non-null   object
 5   All Time Rank               4600 non-null   object
 6   Track Score                 4600 non-null   float64
 7   Spotify Streams             4487 non-null   object
 8   Spotify Playlist Count      4530 non-null   object
 9   Spotify Playlist Reach      4528 non-null   object
 10  Spotify Popularity          3796 non-null   float64
```

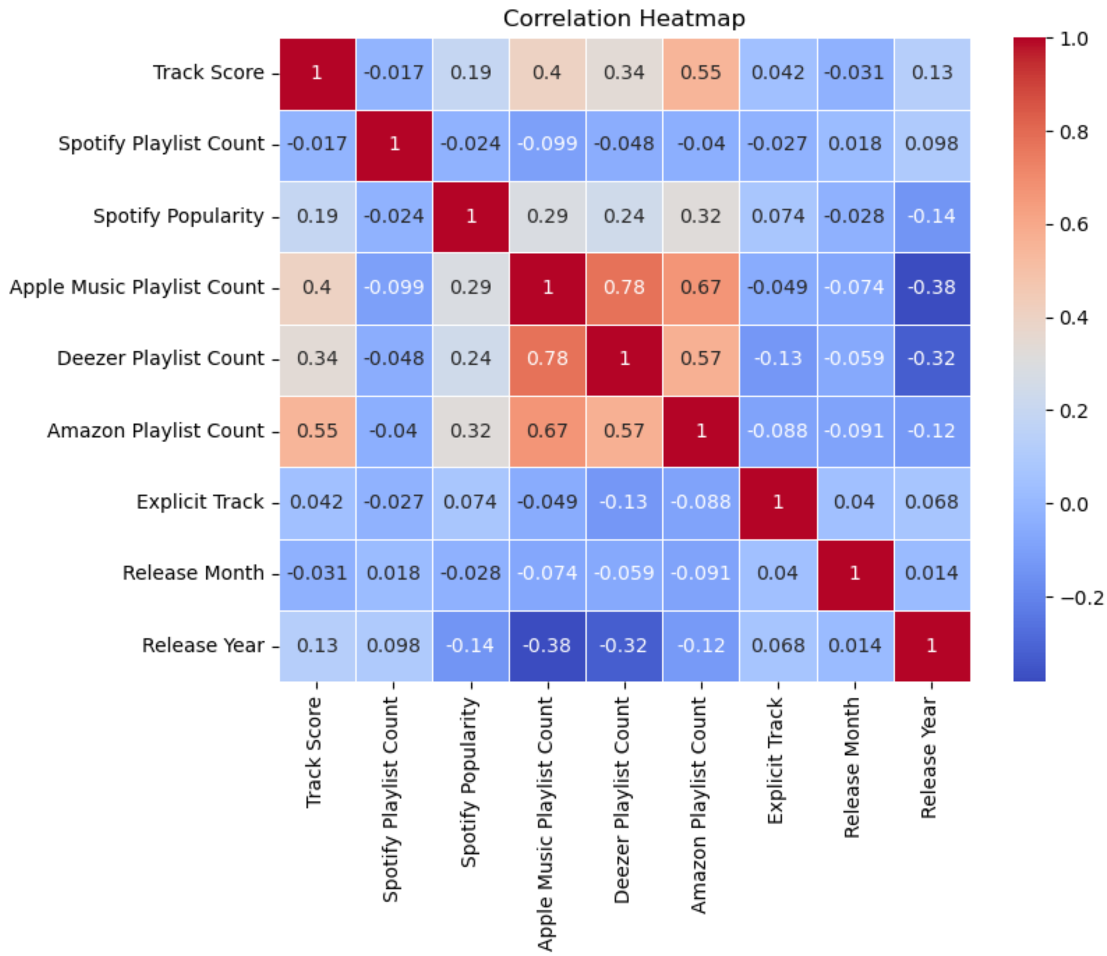
```
 11  YouTube Views               4292 non-null   object
 12  YouTube Likes               4285 non-null   object
 13  TikTok Posts                3427 non-null   object
 14  TikTok Likes                3620 non-null   object
 15  TikTok Views                3619 non-null   object
 16  YouTube Playlist Reach      3591 non-null   object
 17  Apple Music Playlist Count  4039 non-null   float64
 18  AirPlay Spins               4102 non-null   object
 19  SiriusXM Spins              2477 non-null   object
 20  Deezer Playlist Count       3679 non-null   float64
 21  Deezer Playlist Reach       3672 non-null   object
 22  Amazon Playlist Count       3545 non-null   float64
 23  Pandora Streams             3494 non-null   object
 24  Pandora Track Stations      3332 non-null   object
 25  Soundcloud Streams          1267 non-null   object
 26  Shazam Counts               4023 non-null   object
 27  TIDAL Popularity            0 non-null      float64
 28  Explicit Track              4600 non-null   int64
dtypes: float64(6), int64(1), object(22)
memory usage: 1.0+ MB
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 4600 entries, 0 to 4599
Data columns (total 28 columns):
 #   Column                      Non-Null Count  Dtype
---  ------                      --------------  -----
 0   Track                       4600 non-null   object
 1   Album Name                  4600 non-null   object
 2   Artist                      4595 non-null   object
 3   Release Date                4600 non-null   object
 4   ISRC                        4600 non-null   object
 5   All Time Rank               4600 non-null   object
 6   Track Score                 4600 non-null   float64
 7   Spotify Streams             4487 non-null   object
 8   Spotify Playlist Count      4530 non-null   object
 9   Spotify Playlist Reach      4528 non-null   object
 10  Spotify Popularity          3796 non-null   float64
 11  YouTube Views               4292 non-null   object
 12  YouTube Likes               4285 non-null   object
 13  TikTok Posts                3427 non-null   object
 14  TikTok Likes                3620 non-null   object
 15  TikTok Views                3619 non-null   object
 16  YouTube Playlist Reach      3591 non-null   object
 17  Apple Music Playlist Count  4039 non-null   float64
 18  AirPlay Spins               4102 non-null   object
 19  SiriusXM Spins              2477 non-null   object
 20  Deezer Playlist Count       3679 non-null   float64
 21  Deezer Playlist Reach       3672 non-null   object
 22  Amazon Playlist Count       3545 non-null   float64
```

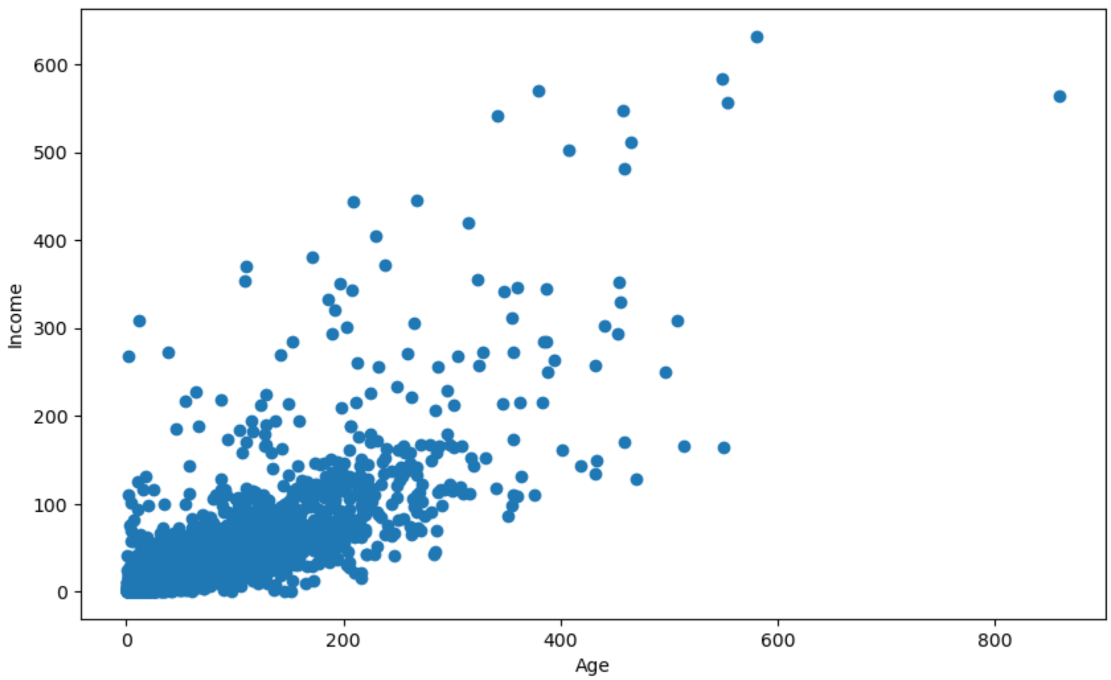
```
 23  Pandora Streams            3494 non-null   object
 24  Pandora Track Stations     3332 non-null   object
 25  Soundcloud Streams         1267 non-null   object
 26  Shazam Counts              4023 non-null   object
 27  Explicit Track             4600 non-null   int64
dtypes: float64(5), int64(1), object(22)
memory usage: 1006.4+ KB
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 4600 entries, 0 to 4599
Data columns (total 28 columns):
 #   Column                      Non-Null Count  Dtype
---  ------                      --------------  -----
 0   Track                       4600 non-null   object
 1   Album Name                  4600 non-null   object
 2   Artist                      4595 non-null   object
 3   Release Date                4600 non-null   object
 4   ISRC                        4600 non-null   object
 5   All Time Rank               4600 non-null   object
 6   Track Score                 4600 non-null   float64
 7   Spotify Streams             4487 non-null   object
 8   Spotify Playlist Count      4600 non-null   int16
 9   Spotify Playlist Reach      4528 non-null   object
 10  Spotify Popularity          3796 non-null   float64
 11  YouTube Views               4292 non-null   object
 12  YouTube Likes               4285 non-null   object
 13  TikTok Posts                3427 non-null   object
 14  TikTok Likes                3620 non-null   object
 15  TikTok Views                3619 non-null   object
 16  YouTube Playlist Reach      3591 non-null   object
 17  Apple Music Playlist Count  4039 non-null   float64
 18  AirPlay Spins               4102 non-null   object
 19  SiriusXM Spins              2477 non-null   object
 20  Deezer Playlist Count       3679 non-null   float64
 21  Deezer Playlist Reach       3672 non-null   object
 22  Amazon Playlist Count       3545 non-null   float64
 23  Pandora Streams             3494 non-null   object
 24  Pandora Track Stations      3332 non-null   object
 25  Soundcloud Streams          1267 non-null   object
 26  Shazam Counts               4023 non-null   object
 27  Explicit Track              4600 non-null   int64
dtypes: float64(5), int16(1), int64(1), object(21)
memory usage: 979.4+ KB
```


Boxplot of Apple Music Playlist Count
Apple Music Playlist Count
800
600
400
200
0
Apple Music Playlist Count


Frequency of Categories
Drake
Taylor Swift
Bad Bunny
KAROL G
The Weeknd
Travis Scott
Billie Eilish
Ariana Grande
Future
Peso Pluma
Post Malone
Juice WRLD
Morgan Wallen
Olivia Rodrigo
Lil Baby
Doja Cat
Rauw Alejandro
Kendrick Lamar
Feid
Dua Lipa
Category
0
10
20
30
40
50
60
Frequency

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 4600 entries, 0 to 4599
Data columns (total 30 columns):
 #   Column                      Non-Null Count  Dtype
---  ------                      --------------  -----
 0   Track                       4600 non-null   object
 1   Album Name                  4600 non-null   object
 2   Artist                      4595 non-null   object
 3   Release Date                4600 non-null   datetime64[ns]
 4   ISRC                        4600 non-null   object
 5   All Time Rank               4600 non-null   object
 6   Track Score                 4600 non-null   float64
 7   Spotify Streams             4487 non-null   object
 8   Spotify Playlist Count      4600 non-null   int16
 9   Spotify Playlist Reach      4528 non-null   object
 10  Spotify Popularity          3796 non-null   float64
 11  YouTube Views               4292 non-null   object
 12  YouTube Likes               4285 non-null   object
 13  TikTok Posts                3427 non-null   object
 14  TikTok Likes                3620 non-null   object
 15  TikTok Views                3619 non-null   object
 16  YouTube Playlist Reach      3591 non-null   object
 17  Apple Music Playlist Count  4039 non-null   float64
 18  AirPlay Spins               4102 non-null   object
 19  SiriusXM Spins              2477 non-null   object
 20  Deezer Playlist Count       3679 non-null   float64
 21  Deezer Playlist Reach       3672 non-null   object
 22  Amazon Playlist Count       3545 non-null   float64
 23  Pandora Streams             3494 non-null   object
 24  Pandora Track Stations      3332 non-null   object
 25  Soundcloud Streams          1267 non-null   object
 26  Shazam Counts               4023 non-null   object
 27  Explicit Track              4600 non-null   int64
 28  Release Month               4600 non-null   int32
 29  Release Year                4600 non-null   int32
dtypes: datetime64[ns](1), float64(5), int16(1), int32(2), int64(1), object(20)
memory usage: 1015.4+ KB
```


Correlation Heatmap
Track Score
Spotify Playlist Count
Spotify Popularity
Apple Music Playlist Count
Deezer Playlist Count
Amazon Playlist Count
Explicit Track
Release Month
Release Year
1.0
0.8
0.6
0.4
0.2
0.0
−0.2



[10]:

Release Year	2015	2016	2017	2018	2019	2020	2021	2022	2023	2024
Release Month										
1	14	18	21	21	19	33	26	32	69	72
2	8	6	17	15	20	38	35	37	88	104
3	6	9	18	16	25	28	29	36	124	133
4	4	12	17	18	20	22	37	48	75	163
5	6	6	16	29	35	27	37	91	82	174
6	11	5	20	21	32	24	41	60	124	47
7	11	9	11	18	21	44	33	57	103	0
8	5	9	25	27	25	25	33	68	91	0
9	6	12	15	17	18	25	42	62	113	0
10	11	12	16	26	27	42	33	69	129	0
11	7	13	21	24	27	34	38	66	90	0
12	7	12	16	17	27	18	24	68	70	0

Conclusion:

Missing Values: The dataset had missing values in the 'TIDAL Popularity' column, which was removed from the analysis. Other numeric features like 'Spotify Playlist Count' had missing values that were imputed with the mean.

Data Types: The 'Spotify Playlist Count' column was converted from a string to an integer data type. Duplicate Records: The dataset had some duplicate records, which were removed, keeping only the first occurrence.

Outliers: Outliers were identified in the 'Track Score' and 'Apple Music Playlist Count' features, but no specific actions were taken to handle them.

Date Formatting: The ‘Release Date’ column was converted to a datetime format, and new columns for ‘Release Month’ and ‘Release Year’ were created.

Univariate Analysis: Numerical variables were analyzed, and the interquartile range (IQR) for ‘Track Score’ was calculated. The top 20 most frequent artists were identified and visualized using a horizontal bar graph.

Bivariate Analysis: A correlation heatmap was created to identify the relationships between the numerical features. A scatter plot was created to visualize the relationship between ‘Apple Music Playlist Count’ and ‘Deezer Playlist Count’.

A cross-tabulation was performed to analyze the relationship between ‘Release Month’ and ‘Release Year’. Insights:

The dataset contains information about the most streamed Spotify songs in 2024, including various attributes such as track score, playlist counts across different platforms, and release date. The data cleaning process revealed some issues with missing values, data types, and duplicate records, which were addressed accordingly. The univariate analysis provided insights into the distribution and popularity of different artists, with the top 20 artists being identified. The bivariate analysis revealed some interesting relationships between the numerical features. The correlation heatmap showed that ‘Apple Music Playlist Count’ and ‘Deezer Playlist Count’ are highly correlated, which was further confirmed by the scatter plot. The cross-tabulation of ‘Release Month’ and ‘Release Year’ could provide insights into the seasonality and trends in the release of the most streamed Spotify songs over the years. These insights could help inform decision-making in the music industry, such as:

Identifying popular artists and genres to focus on for future releases or promotional efforts. Understanding the relationship between playlist counts across different platforms, which could inform cross-platform marketing strategies. Identifying seasonal trends in song releases, which could guide release timing and planning. Overall, the analysis of this dataset provides a comprehensive understanding of the most streamed Spotify songs in 2024 and could be leveraged to make informed decisions in the music industry.