**Week 2 Assignment**

Resha Maharjan

Department of IT, Westcliff University

Presidential Graduate School

TECH 405: Artificial Neural Network & Deep Learning

Professor Acharya

November 4, 2024

Table of Contents

[Introduction 2](#_Toc181697950)

[Findings 2](#_Toc181697951)

[Approach 3](#_Toc181697952)

[Conclusion 3](#_Toc181697953)

# **Introduction**

This report aims to explore the trends and association in the music dataset by analyzing the “spotify\_dataset.csv”. I have used Python packages to read, preprocess and analyze the data for visualization purposes. I find out how songs are distributed and examine possible patterns of such features as top artists, top genres, charting positions, number of times charted, and overall popularity using histograms, count plots, correlation matrices, and pair plots.

# **Findings**

The findings form the EDA that we got from “spotify\_dataset.csv” dataset are:

1. **Popularity Distribution:** The shape of the histogram constructed for the “Popularity” column show’s how it is skewed high or low, among other things.
2. **Top Artists, Genres, and Song Names:** The frequency distribution of plots of the “Artist”, “Song Name” and “Genre” know the most frequently used are the artist, the genre and even the song names that may be reproduced.
3. **Correlations between Charting and Popularity:** The relative correlation between “Highest Charting Position”, “Number of Times Charted”, and “Popularity” will be easily determined from the correlation matrix. A positive correlation indicates that both number of times charted and chart position are both positive and consequently popularity.
4. **Outliers in Charting and Popularity:** Removing outliers for numerical variables can also be done, the outliers can be identified by creating a boxplot for “Highest Charting Position”, “Number of times charted’ and ‘Popularity”. These could be better discussed in another section.

# **Approach**

The challenges in the dataset through several approaches:

1. **Handling Missing Values:** I first started with ensuring that there were no missing values for instance, I used functions like ‘df.info()’ and or ‘df.isnull().sum()’. This enabled me to see the gaps in data with reference to the data set at my disposal.
2. **Data Type Conversion:** In order to make all the data types of the column coherent with each other I checked the existing data types of all the columns using the code ‘df.dtypes’. This let me further modify as needed, especially for fields such as “Popularity,” which needed to be numerical for further analysis. I also used box plots to console the two columns such as “Highest Charting Position” and the “Number of Times Charted”
3. **Outlier Detection:** For the columns such as “Highest Charting Position” and “Number of Times Charted”, the code draws boxplots to capture any outlying data that might distort analysis.

**Conclusion**

The code also encompasses the exploratory data analysis commonly referred to as the EDA on a Spotify dataset.et. Data Pre-processing process starts with data loading and cleaning and missing values treatment and getting to know the distribution of the songs popularity. The analysis then moves out to determine the best artists, genres and song names, to uncover the most successful categories within the set. To do this the code computes correlation matrix and a pair plot to show the nature of the correlations between the numerical features. Besides, they help to identify possible outliers that might affect the entire analysis.Toed plots can be used to explain the distribution of a particular song most popularity, top shaped artist and genre, as well as relationships between chart performances and overall most popularity. It also identifies thresholds of numerical features, which gives a general idea about the existence of outliers in context of general trends of the data that is available in Spotify dataset.

# **Reference**

Courtney, M. B. (2021). Exploratory Data Analysis in Schools: A Logic Model to Guide Implementation. *International Journal of Education Policy and Leadership*, *17*(4). <https://doi.org/10.22230/ijepl.2021v17n4a1041>

Ho Yu, C. (2010). Exploratory data analysis in the context of data mining and resampling. *International Journal of Psychological Research*, *3*(1), 9–22. <https://doi.org/10.21500/20112084.819>

Kaggle. (2019). *Kaggle: Your Home for Data Science*. Kaggle.com. <https://www.kaggle.com>

‌

‌