TECHNOLOGICAL UNIVERSITY OF THE PHILIPPINES – MANILA
College of Engineering
Electronics Engineering Department
Ayala Blvd., Ermita, Manila, 1000, Philippines
Tel. No. +632-301-3001 | Fax No. +632-521-4063

Programming Assignment 1

EXPLORATORY DATA ANALYSIS

In this assignment, you will perform an exploratory data analysis (EDA) on a dataset containing information about popular tracks on **Most Streamed Spotify Songs 2023** (https://www.kaggle.com/datasets/nelgiriyewithana/top-spotify-songs-2023). The goal of this task is to analyze, visualize, and interpret the data to extract meaningful insights.

General Guidelines

1. Begin by familiarizing yourself with the structure of the dataset. Check for missing values, data types, and perform an initial exploration to understand the different features available.
2. Provide summary statistics to give an overview of key metrics such as the number of streams, release dates, and musical attributes (e.g., BPM, danceability).
3. Use appropriate visualizations (e.g., bar charts, histograms, scatter plots) to uncover trends and patterns in the data. Ensure that your plots are well-labeled and easy to interpret.
4. Investigate correlations between different variables and provide insights based on your findings. Explore relationships between streams and other musical characteristics like tempo, energy, or playlists.
5. Based on your analysis, offer any insights or recommendations regarding the tracks, artists, or musical trends that could be useful for understanding what makes a track popular.

Guide Questions

You are expected to answer the following questions using your analysis:

1. **Overview of Dataset**
   - How many rows and columns does the dataset contain?
   - What are the data types of each column? Are there any missing values?
2. **Basic Descriptive Statistics**
   - What are the mean, median, and standard deviation of the `streams` column?
   - What is the distribution of `released_year` and `artist_count`? Are there any noticeable trends or outliers?
3. **Top Performers**
   - Which track has the highest number of `streams`? Display the top 5 most streamed tracks.
   - Who are the top 5 most frequent artists based on the number of tracks in the dataset?
4. **Temporal Trends**
   - Analyze the trends in the number of tracks released over time. Plot the number of tracks released per year.
   - Does the number of tracks released per month follow any noticeable patterns? Which month sees the most releases?
5. **Genre and Music Characteristics**
   - Examine the correlation between `streams` and musical attributes like `bpm`, `danceability_%`, and `energy_%`. Which attributes seem to influence streams the most?
   - Is there a correlation between `danceability_%` and `energy_%`? How about `valence_%` and `acousticness_%`?

6. **Platform Popularity**
   - How do the numbers of tracks in `spotify_playlists`, `spotify_charts`, and `apple_playlists` compare? Which platform seems to favor the most popular tracks?
7. **Advanced Analysis**
   - Based on the `streams` data, can you identify any patterns among tracks with the same key or mode (Major vs. Minor)?
   - Do certain genres or artists consistently appear in more playlists or charts? Perform an analysis to compare the most frequently appearing artists in playlists or charts.

Requirements

- Ensure that your code is clean, well-commented, and organized.
- Use Python libraries such as `pandas` for data manipulation and `matplotlib` or `seaborn` for visualization.
- Provide a brief written interpretation of each key insight you discover.

Submission

1. Submit your work as a Jupyter Notebook (`.ipynb`) file.
2. Upload your Jupyter Notebook to your GitHub repository. Ensure the notebook is well-documented with markdown cells explaining each step and the corresponding results.
3. Provide the link to your GitHub repository for grading.