HW4

Dataset Overview

For this exercise, we will use a dataset containing information about the most streamed songs on Spotify. The dataset includes columns like `track_name`, `artist(s)_name`, `streams`, `danceability_percent`, `valence_percent`, and other musical features. Your task is to perform data analysis and manipulation using Python's `pandas` library.

Exercise 1: Basic Data Inspection

1. Load the dataset and display the first and last five rows.
2. Check the dataset shape (rows, columns) and the data types of each column.
3. List all column names in the dataset.
4. Find and count any missing values in the dataset.

Exercise 2: Filtering and Sorting Data

1. Filter the dataset for songs released in the year 2023.
2. Sort the dataset by the `streams` column in descending order to find the most streamed songs.
3. Filter the songs that have a `danceability_percent` greater than 70 and an `energy_percent` greater than 80.

Exercise 3: GroupBy and Aggregation

1. Group the dataset by `artist` and calculate the total number of streams for each artist.
2. Find the artist with the most tracks in the dataset.
3. Calculate the average `danceability_percent`, `valence_percent`, and `energy_percent` for songs released in each year.

Exercise 4: Manipulating Columns

1. Create a new column that calculates the total number of playlists a song appears on (sum of `in_spotify_playlists` and `in_apple_playlists`).
2. Normalize the `streams` column by applying min-max scaling.
3. Convert the `released_year`, `released_month`, and `released_day` columns into a single `release_date` column of type `datetime`.

Exercise 5: Descriptive Statistics

1. Calculate the mean, median, and standard deviation for the `streams`, `danceability_percent`, and `energy_percent` columns.
2. Find the song with the highest and lowest `valence_percent`.

Exercise 6: Indexing and Slicing

1. Select specific columns (e.g., `track_name`, `artist(s)_name`, `streams`) from the dataframe.
2. Use `.loc` and `.iloc` to slice the dataframe and display the first 10 rows with their corresponding `track_name` and `streams`.