TASK REPORT

ON

Exploratory Data Analysis (EDA)

Intern Name: Ganesh Wagh

Internship Domain: Data Analysis

Organization: Codealpha

Task Name : Task 2 - Exploratory Data Analysis (EDA)

Tools Used : Python (Pandas, Matplotlib) using Jupyter Notebook

Duration: 10 January 2026 – 24 February 2026

Abstract

Exploratory Data Analysis (EDA) is a crucial step in data analysis that helps in understanding the structure, characteristics, and quality of a dataset. In this task, EDA was performed on the Spotify Tracks dataset to analyze its features and gain meaningful insights. The analysis involved examining the dataset structure, identifying data types, checking for missing values and duplicate records, and studying statistical summaries of numerical attributes. Data visualization techniques such as histograms, correlation analysis, and box plots were used to identify patterns, relationships, and outliers within the data. The results of this analysis provided a clear understanding of the dataset and highlighted potential data issues that may require preprocessing before further analysis. This task demonstrates the importance of EDA in preparing data for informed decision-making and advanced analytical tasks.

Introduction

Exploratory Data Analysis (EDA) is an important step in data analysis that helps in understanding the structure, characteristics, and quality of the dataset. It involves summarizing the data, identifying patterns, trends, relationships, and detecting anomalies or data issues before further analysis.

Dataset Description

The dataset used for this task is the Spotify Tracks Dataset, obtained from Kaggle.

The dataset contains information about Spotify songs along with their audio features such as popularity, danceability, energy, tempo, and other attributes.

Objectives

- To understand the structure of the dataset
- To identify data types and missing values
- To detect duplicate records
- To analyze distributions of numerical features
- To study relationships between audio features
- To identify outliers and anomalies

Questions Asked Before Analysis

- How many records and features are present in the dataset?
- What types of variables are included?
- Are there missing or duplicate values?
- How are audio features like popularity, energy, and danceability distributed?
- Is there any relationship between popularity and other audio features?
- Are there outliers present in the data?

Data Structure and Overview

The dataset consists of multiple rows representing individual Spotify tracks.
Columns include both categorical variables (track name, artist name) and numerical variables (popularity, danceability, energy, tempo).
Data types were examined to differentiate between numerical and textual data.

Summary Statistics

Descriptive statistics such as mean, minimum, maximum, and standard deviation were used to understand the central tendency and spread of numerical features. This helped in identifying the general range and variation of audio attributes

Missing Values and Duplicates

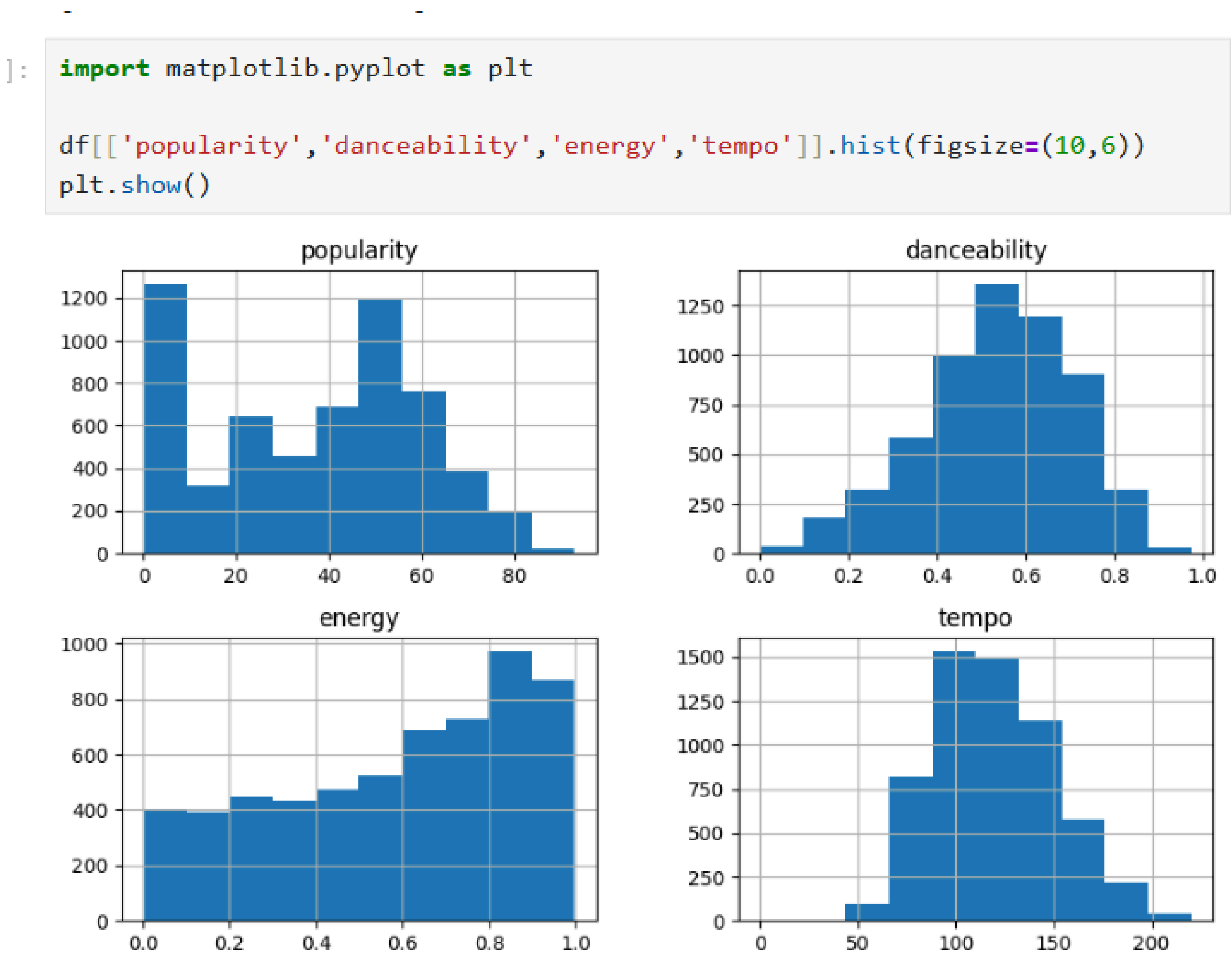
Missing values were checked across all columns.
Duplicate records were analyzed, and the dataset was found to have no duplicate rows, indicating good data quality.

Distribution Analysis

Histograms were used to analyze the distribution of important numerical features such as:

- Popularity
- Danceability
- Energy
- Tempo

The distributions revealed how frequently different values occur and helped identify skewness in some features.



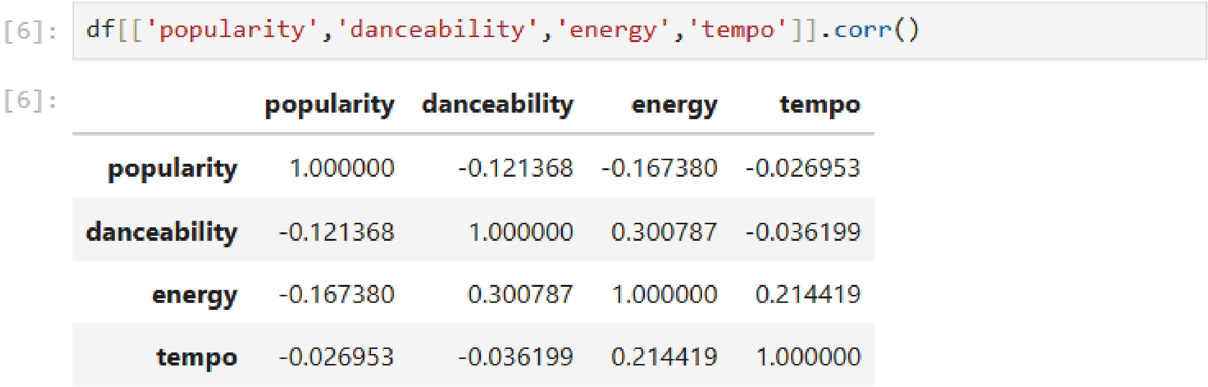
```
import matplotlib.pyplot as plt

df[['popularity','danceability','energy','tempo']].hist(figsize=(10,6))
plt.show()
```

Figure 1: Distribution of Spotify audio features

Relationship Analysis

A correlation analysis was performed to understand relationships between numerical variables.

The correlation matrix helped in identifying whether features like energy, danceability, and tempo have any influence on track popularity.



```
[6]: df[['popularity','danceability','energy','tempo']].corr()
```

[6]:

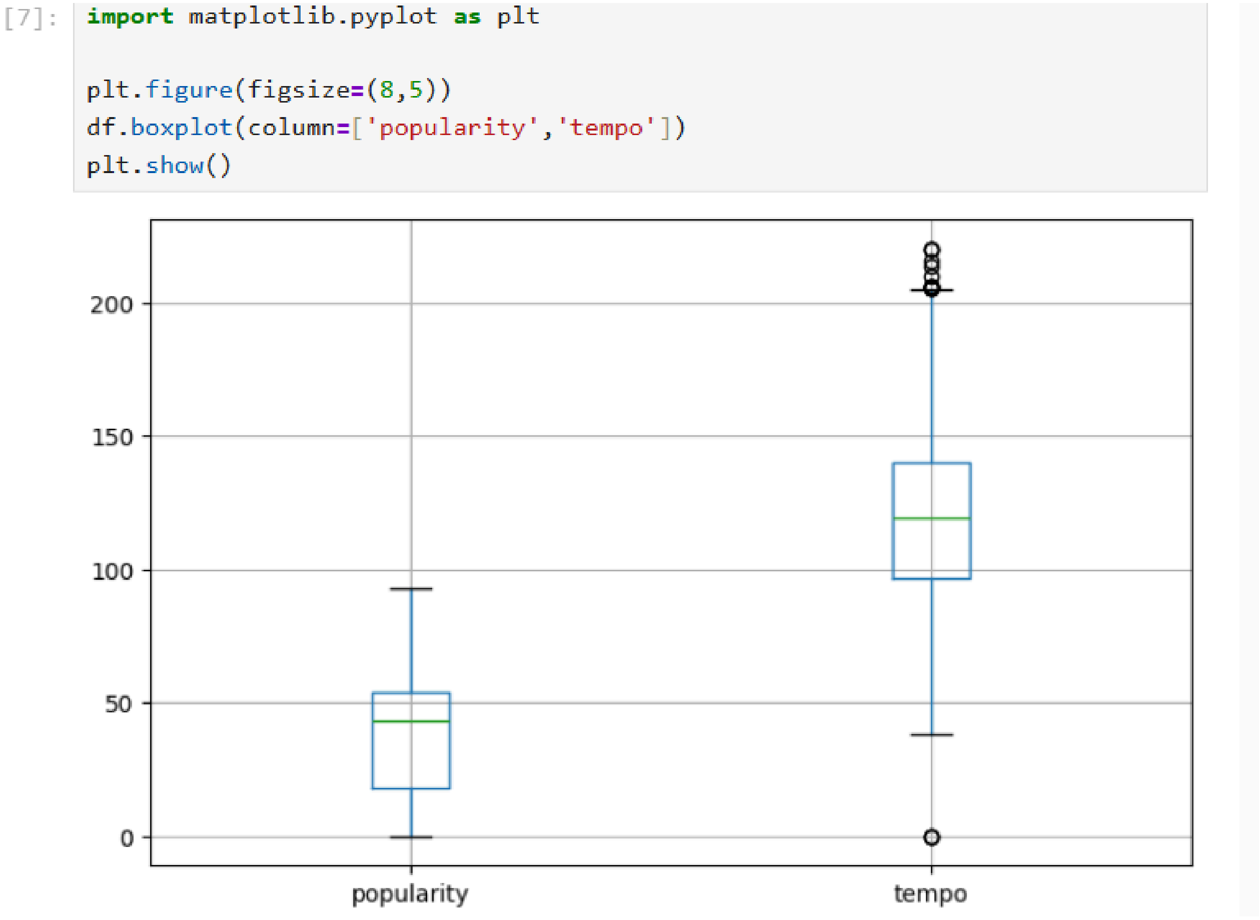
	popularity	danceability	energy	tempo
popularity	1.000000	-0.121368	-0.167380	-0.026953
danceability	-0.121368	1.000000	0.300787	-0.036199
energy	-0.167380	0.300787	1.000000	0.214419
tempo	-0.026953	-0.036199	0.214419	1.000000

Figure 2: Correlation matrix of audio features

Outlier Detection

Box plots were used to detect outliers in numerical columns such as popularity and tempo.

Outliers indicate unusually high or low values, which may represent rare cases or require further investigation.



```
[7]: import matplotlib.pyplot as plt

     plt.figure(figsize=(8,5))
     df.boxplot(column=['popularity','tempo'])
     plt.show()
```

Figure 3: Box plot showing outliers in popularity and tempo

Data Issues Identified

- Some numerical features show skewed distributions
- Presence of outliers in tempo and popularity
- Dataset may require preprocessing before advanced analysis

Conclusion

Exploratory Data Analysis (EDA) was successfully performed on the Spotify Tracks dataset to understand its structure, distribution, and key characteristics. Statistical analysis and visualizations helped identify patterns, relationships among audio features, and the presence of outliers. The analysis provided valuable insights into the dataset and highlighted potential data quality issues, emphasizing the importance of preprocessing before further analysis.

Future Scope

- Handling outliers and skewed distributions
- Applying feature engineering techniques
- Building predictive models for track popularity
- Performing advanced data visualization

Outcome of the Task

Gained understanding of dataset structure and features
Identified data quality issues such as outliers and skewness
Analyzed relationships between numerical variables
Prepared the dataset for further analysis