Task-02

Perform data cleaning and exploratory data analysis (EDA) on a dataset of your choice, such as the Titanic dataset from Kaggle. Explore the relationships between variables and identify patterns and trends in the data.

Sample Dataset :- [https://www.kaggle.com/c/titanic/data](https://www.google.com/url?q=https%3A%2F%2Fwww.kaggle.com%2Fc%2Ftitanic%2Fdata)

# Import necessary libraries

import pandas as pd

import numpy as np

import matplotlib.pyplot as plt

import seaborn as sns

df=pd.read\_csv('iris.csv')

df.head()

print("Dataset Info:")

print(df.info())

print("\nSummary Statistics:")

print(df.describe())

print("\nChecking for missing values:")

print(df.isnull().sum())

print("\nDuplicates:", df.duplicated().sum())

# Drop duplicates if any

df.drop\_duplicates(inplace=True)

sns.countplot(x='variety', data=df)

plt.title("Species Distribution")

plt.show()

sns.pairplot(df, hue='variety', markers=["o", "s", "D"])

plt.suptitle("Pairwise Plot of Features", y=1.02)

plt.show()

plt.figure(figsize=(12, 8))

for idx, feature in enumerate(df.columns[:-1]):

    plt.subplot(2, 2, idx+1)

    sns.boxplot(x='variety', y=feature, data=df)

    plt.title(f"{feature.capitalize()} by Species")

plt.tight\_layout()

plt.show()

plt.figure(figsize=(8, 6))

sns.heatmap(df.iloc[:, :-1].corr(), annot=True, cmap='coolwarm')

plt.title("Feature Correlation Matrix")

plt.show()

plt.figure(figsize=(12, 8))

for idx, feature in enumerate(df.columns[:-1]):

    plt.subplot(2, 2, idx+1)

    sns.violinplot(x='variety', y=feature, data=df)

    plt.title(f"Distribution of {feature} by Species")

plt.tight\_layout()

plt.show()

df.iloc[:, :-2].hist(figsize=(10, 6))

plt.suptitle("Histograms of Numeric Features")

plt.show()

sns.histplot(data=df, x='petal.length', bins=20, kde=False, color='red')

plt.title('Histogram of Petal Length (No KDE)')

plt.xlabel('Petal Length (cm)')

plt.ylabel('Frequency')

plt.show()

sns.histplot(data=df, x='petal.length', bins=20, kde=True, color='blue')

plt.title('Histogram with KDE of Petal Length')

plt.xlabel('Petal Length (cm)')

plt.ylabel('Density')

plt.show()

sns.kdeplot(data=df, x='petal.length', fill=True, color='green')

plt.title('KDE Plot of Petal Length (No Histogram)')

plt.xlabel('Petal Length')

plt.ylabel('Density')

plt.show()