import pandas as pd

import numpy as np

import matplotlib.pyplot as plt

import seaborn as sns

# Load the dataset

dataset\_path = '/mnt/data/github\_dataset.csv'

df1 = pd.read\_csv(dataset\_path)

# Display basic information

print("Dataset Shape:", df1.shape)

print("\nColumn Information:")

print(df1.info())

# Display first few rows

print("\nFirst 5 Rows:")

print(df1.head())

# Check for missing values

print("\nMissing Values:")

print(df1.isnull().sum())

# Check for duplicates

duplicates = df1.duplicated().sum()

print("\nDuplicate Rows:", duplicates)

# Summary statistics for numerical columns

print("\nNumerical Summary:")

print(df1.describe())

# Summary statistics for categorical columns

print("\nCategorical Summary:")

categorical\_cols = df1.select\_dtypes(include=['object']).columns

for col in categorical\_cols:

print(f"{col}: {df1[col].nunique()} unique values")

print(df1[col].value\_counts().head())

print()

# Visualization - Histograms

df1.hist(figsize=(12, 8), bins=30)

plt.suptitle("Feature Distributions", fontsize=16)

plt.show()

# Correlation Heatmap

plt.figure(figsize=(10, 6))

sns.heatmap(df1.corr(), annot=True, cmap='coolwarm', fmt='.2f')

plt.title("Feature Correlation Heatmap")

plt.show()