import sys

import pandas as pd

# Check if the data location argument is provided

if len(sys.argv) != 2:

print("Usage: python data\_processing.py <data\_location>")

sys.exit(1)

# Load the raw dataset (step 1)

df = pd.read\_csv(sys.argv[1])

# Rename columns to more descriptive names (step 2)

df.columns = [

"Country",

"Happiness Score",

"Happiness Score Error",

"Upper Whisker",

"Lower Whisker",

"GDP per Capita",

"Social Support",

"Healthy Life Expectancy",

"Freedom to Make Life Choices",

"Generosity",

"Perceptions of Corruption",

"Dystopia Happiness Score",

"GDP per Capita",

"Social Support",

"Healthy Life Expectancy",

"Freedom to Make Life Choices",

"Generosity",

"Perceptions of Corruption",

"Dystopia Residual",

]

# Handle missing values by replacing them with the mean (step 3)

df.fillna(df.mean(numeric\_only=True), inplace=True)

# Check for missing values after cleaning

print("Missing values after cleaning:")

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

print(df.head())

# Save the cleaned and normalized dataset to a new CSV file (step 4)

df.to\_csv("processed\_data\WHR2023\_cleaned.csv", index=False)