EXPERIMENT 5

# Import necessary libraries

import pandas as pd

import numpy as np

import matplotlib.pyplot as plt

import seaborn as sns

# Load the dataset

dataset = pd.read\_csv('/kaggle/input/life-expectancy-who/Life Expectancy Data.csv')

# Display first few rows of the dataset

print("First 5 rows of the dataset:")

print(dataset.head())

First 5 rows of the dataset:

Country Year Status Life expectancy Adult Mortality \

0 Afghanistan 2015 Developing 65.0 263.0

1 Afghanistan 2014 Developing 59.9 271.0

2 Afghanistan 2013 Developing 59.9 268.0

3 Afghanistan 2012 Developing 59.5 272.0

4 Afghanistan 2011 Developing 59.2 275.0

infant deaths Alcohol percentage expenditure Hepatitis B Measles ... \

0 62 0.01 71.279624 65.0 1154 ...

1 64 0.01 73.523582 62.0 492 ...

2 66 0.01 73.219243 64.0 430 ...

3 69 0.01 78.184215 67.0 2787 ...

4 71 0.01 7.097109 68.0 3013 ...

Polio Total expenditure Diphtheria HIV/AIDS GDP Population \

0 6.0 8.16 65.0 0.1 584.259210 33736494.0

1 58.0 8.18 62.0 0.1 612.696514 327582.0

2 62.0 8.13 64.0 0.1 631.744976 31731688.0

3 67.0 8.52 67.0 0.1 669.959000 3696958.0

4 68.0 7.87 68.0 0.1 63.537231 2978599.0

thinness 1-19 years thinness 5-9 years \

0 17.2 17.3

1 17.5 17.5

2 17.7 17.7

3 17.9 18.0

4 18.2 18.2

Income composition of resources Schooling

0 0.479 10.1

1 0.476 10.0

2 0.470 9.9

3 0.463 9.8

4 0.454 9.5

[5 rows x 22 columns]

# Basic information about the dataset

print("\nDataset Info:")

print(dataset.info())

Dataset Info:

<class 'pandas.core.frame.DataFrame'>

RangeIndex: 2938 entries, 0 to 2937

Data columns (total 22 columns):

# Column Non-Null Count Dtype

--- ------ -------------- -----

0 Country 2938 non-null object

1 Year 2938 non-null int64

2 Status 2938 non-null object

3 Life expectancy 2928 non-null float64

4 Adult Mortality 2928 non-null float64

5 infant deaths 2938 non-null int64

6 Alcohol 2744 non-null float64

7 percentage expenditure 2938 non-null float64

8 Hepatitis B 2385 non-null float64

9 Measles 2938 non-null int64

10 BMI 2904 non-null float64

11 under-five deaths 2938 non-null int64

12 Polio 2919 non-null float64

13 Total expenditure 2712 non-null float64

14 Diphtheria 2919 non-null float64

15 HIV/AIDS 2938 non-null float64

16 GDP 2490 non-null float64

17 Population 2286 non-null float64

18 thinness 1-19 years 2904 non-null float64

19 thinness 5-9 years 2904 non-null float64

20 Income composition of resources 2771 non-null float64

21 Schooling 2775 non-null float64

dtypes: float64(16), int64(4), object(2)

memory usage: 505.1+ KB

None

# Statistical summary of numerical columns

print("\nStatistical Summary:")

print(dataset.describe())

Statistical Summary:

Year Life expectancy Adult Mortality infant deaths \

count 2938.000000 2928.000000 2928.000000 2938.000000

mean 2007.518720 69.224932 164.796448 30.303948

std 4.613841 9.523867 124.292079 117.926501

min 2000.000000 36.300000 1.000000 0.000000

25% 2004.000000 63.100000 74.000000 0.000000

50% 2008.000000 72.100000 144.000000 3.000000

75% 2012.000000 75.700000 228.000000 22.000000

max 2015.000000 89.000000 723.000000 1800.000000

Alcohol percentage expenditure Hepatitis B Measles \

count 2744.000000 2938.000000 2385.000000 2938.000000

mean 4.602861 738.251295 80.940461 2419.592240

std 4.052413 1987.914858 25.070016 11467.272489

min 0.010000 0.000000 1.000000 0.000000

25% 0.877500 4.685343 77.000000 0.000000

50% 3.755000 64.912906 92.000000 17.000000

75% 7.702500 441.534144 97.000000 360.250000

max 17.870000 19479.911610 99.000000 212183.000000

BMI under-five deaths Polio Total expenditure \

count 2904.000000 2938.000000 2919.000000 2712.00000

mean 38.321247 42.035739 82.550188 5.93819

std 20.044034 160.445548 23.428046 2.49832

min 1.000000 0.000000 3.000000 0.37000

25% 19.300000 0.000000 78.000000 4.26000

50% 43.500000 4.000000 93.000000 5.75500

75% 56.200000 28.000000 97.000000 7.49250

max 87.300000 2500.000000 99.000000 17.60000

Diphtheria HIV/AIDS GDP Population \

count 2919.000000 2938.000000 2490.000000 2.286000e+03

mean 82.324084 1.742103 7483.158469 1.275338e+07

std 23.716912 5.077785 14270.169342 6.101210e+07

min 2.000000 0.100000 1.681350 3.400000e+01

25% 78.000000 0.100000 463.935626 1.957932e+05

50% 93.000000 0.100000 1766.947595 1.386542e+06

75% 97.000000 0.800000 5910.806335 7.420359e+06

max 99.000000 50.600000 119172.741800 1.293859e+09

thinness 1-19 years thinness 5-9 years \

count 2904.000000 2904.000000

mean 4.839704 4.870317

std 4.420195 4.508882

min 0.100000 0.100000

25% 1.600000 1.500000

50% 3.300000 3.300000

75% 7.200000 7.200000

max 27.700000 28.600000

Income composition of resources Schooling

count 2771.000000 2775.000000

mean 0.627551 11.992793

std 0.210904 3.358920

min 0.000000 0.000000

25% 0.493000 10.100000

50% 0.677000 12.300000

75% 0.779000 14.300000

max 0.948000 20.700000

# Check for missing values

print("\nMissing Values:")

print(dataset.isnull().sum())

Missing Values:

Country 0

Year 0

Status 0

Life expectancy 10

Adult Mortality 10

infant deaths 0

Alcohol 194

percentage expenditure 0

Hepatitis B 553

Measles 0

BMI 34

under-five deaths 0

Polio 19

Total expenditure 226

Diphtheria 19

HIV/AIDS 0

GDP 448

Population 652

thinness 1-19 years 34

thinness 5-9 years 34

Income composition of resources 167

Schooling 163

dtype: int64

# Distribution of numerical columns

numerical\_columns = dataset.select\_dtypes(include=[np.number]).columns.tolist()

for col in numerical\_columns:

plt.figure(figsize=(6, 4))

sns.histplot(dataset[col], kde=True, bins=30)

plt.title(f"Distribution of {col}")

plt.show()

/opt/conda/lib/python3.10/site-packages/seaborn/\_oldcore.py:1119: FutureWarning: use\_inf\_as\_na option is deprecated and will be removed in a future version. Convert inf values to NaN before operating instead.

with pd.option\_context('mode.use\_inf\_as\_na', True):

/opt/conda/lib/python3.10/site-packages/seaborn/\_oldcore.py:1119: FutureWarning: use\_inf\_as\_na option is deprecated and will be removed in a future version. Convert inf values to NaN before operating instead.

with pd.option\_context('mode.use\_inf\_as\_na', True):

/opt/conda/lib/python3.10/site-packages/seaborn/\_oldcore.py:1119: FutureWarning: use\_inf\_as\_na option is deprecated and will be removed in a future version. Convert inf values to NaN before operating instead.

with pd.option\_context('mode.use\_inf\_as\_na', True):

/opt/conda/lib/python3.10/site-packages/seaborn/\_oldcore.py:1119: FutureWarning: use\_inf\_as\_na option is deprecated and will be removed in a future version. Convert inf values to NaN before operating instead.

with pd.option\_context('mode.use\_inf\_as\_na', True):

/opt/conda/lib/python3.10/site-packages/seaborn/\_oldcore.py:1119: FutureWarning: use\_inf\_as\_na option is deprecated and will be removed in a future version. Convert inf values to NaN before operating instead.

with pd.option\_context('mode.use\_inf\_as\_na', True):

/opt/conda/lib/python3.10/site-packages/seaborn/\_oldcore.py:1119: FutureWarning: use\_inf\_as\_na option is deprecated and will be removed in a future version. Convert inf values to NaN before operating instead.

with pd.option\_context('mode.use\_inf\_as\_na', True):

/opt/conda/lib/python3.10/site-packages/seaborn/\_oldcore.py:1119: FutureWarning: use\_inf\_as\_na option is deprecated and will be removed in a future version. Convert inf values to NaN before operating instead.

with pd.option\_context('mode.use\_inf\_as\_na', True):

/opt/conda/lib/python3.10/site-packages/seaborn/\_oldcore.py:1119: FutureWarning: use\_inf\_as\_na option is deprecated and will be removed in a future version. Convert inf values to NaN before operating instead.

with pd.option\_context('mode.use\_inf\_as\_na', True):

/opt/conda/lib/python3.10/site-packages/seaborn/\_oldcore.py:1119: FutureWarning: use\_inf\_as\_na option is deprecated and will be removed in a future version. Convert inf values to NaN before operating instead.

with pd.option\_context('mode.use\_inf\_as\_na', True):

/opt/conda/lib/python3.10/site-packages/seaborn/\_oldcore.py:1119: FutureWarning: use\_inf\_as\_na option is deprecated and will be removed in a future version. Convert inf values to NaN before operating instead.

with pd.option\_context('mode.use\_inf\_as\_na', True):

/opt/conda/lib/python3.10/site-packages/seaborn/\_oldcore.py:1119: FutureWarning: use\_inf\_as\_na option is deprecated and will be removed in a future version. Convert inf values to NaN before operating instead.

with pd.option\_context('mode.use\_inf\_as\_na', True):

/opt/conda/lib/python3.10/site-packages/seaborn/\_oldcore.py:1119: FutureWarning: use\_inf\_as\_na option is deprecated and will be removed in a future version. Convert inf values to NaN before operating instead.

with pd.option\_context('mode.use\_inf\_as\_na', True):

/opt/conda/lib/python3.10/site-packages/seaborn/\_oldcore.py:1119: FutureWarning: use\_inf\_as\_na option is deprecated and will be removed in a future version. Convert inf values to NaN before operating instead.

with pd.option\_context('mode.use\_inf\_as\_na', True):

/opt/conda/lib/python3.10/site-packages/seaborn/\_oldcore.py:1119: FutureWarning: use\_inf\_as\_na option is deprecated and will be removed in a future version. Convert inf values to NaN before operating instead.

with pd.option\_context('mode.use\_inf\_as\_na', True):

/opt/conda/lib/python3.10/site-packages/seaborn/\_oldcore.py:1119: FutureWarning: use\_inf\_as\_na option is deprecated and will be removed in a future version. Convert inf values to NaN before operating instead.

with pd.option\_context('mode.use\_inf\_as\_na', True):

/opt/conda/lib/python3.10/site-packages/seaborn/\_oldcore.py:1119: FutureWarning: use\_inf\_as\_na option is deprecated and will be removed in a future version. Convert inf values to NaN before operating instead.

with pd.option\_context('mode.use\_inf\_as\_na', True):

/opt/conda/lib/python3.10/site-packages/seaborn/\_oldcore.py:1119: FutureWarning: use\_inf\_as\_na option is deprecated and will be removed in a future version. Convert inf values to NaN before operating instead.

with pd.option\_context('mode.use\_inf\_as\_na', True):

/opt/conda/lib/python3.10/site-packages/seaborn/\_oldcore.py:1119: FutureWarning: use\_inf\_as\_na option is deprecated and will be removed in a future version. Convert inf values to NaN before operating instead.

with pd.option\_context('mode.use\_inf\_as\_na', True):

/opt/conda/lib/python3.10/site-packages/seaborn/\_oldcore.py:1119: FutureWarning: use\_inf\_as\_na option is deprecated and will be removed in a future version. Convert inf values to NaN before operating instead.

with pd.option\_context('mode.use\_inf\_as\_na', True):

/opt/conda/lib/python3.10/site-packages/seaborn/\_oldcore.py:1119: FutureWarning: use\_inf\_as\_na option is deprecated and will be removed in a future version. Convert inf values to NaN before operating instead.

with pd.option\_context('mode.use\_inf\_as\_na', True):

# Select only numerical columns for correlation matrix

numerical\_dataset = dataset.select\_dtypes(include=[np.number])

# Check if there are numerical columns before plotting

if numerical\_dataset.shape[1] > 0:

# Compute correlation matrix

correlation\_matrix = numerical\_dataset.corr()

# Plot heatmap

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

sns.heatmap(correlation\_matrix, annot=True, cmap='coolwarm', fmt='.2f')

plt.title("Correlation Heatmap")

plt.show()

else:

print("No numerical columns available for correlation.")

# Boxplot to check for outliers

for col in numerical\_columns:

plt.figure(figsize=(6, 4))

sns.boxplot(data=df, x=col)

plt.title(f"Boxplot of {col}")

plt.show()

# Convert all numerical columns to the correct type (if necessary)

numerical\_dataset = dataset.select\_dtypes(include=[np.number])

dataset[numerical\_dataset.columns] = numerical\_dataset.apply(pd.to\_numeric, errors='coerce')

# Fill missing values with the median for numerical columns

dataset[numerical\_dataset.columns] = dataset[numerical\_dataset.columns].fillna(dataset[numerical\_dataset.columns].median())

# Re-check missing values

print("\nMissing Values after filling:")

print(dataset.isnull().sum())

Missing Values after filling:

Country 0

Year 0

Status 0

Life expectancy 0

Adult Mortality 0

infant deaths 0

Alcohol 0

percentage expenditure 0

Hepatitis B 0

Measles 0

BMI 0

under-five deaths 0

Polio 0

Total expenditure 0

Diphtheria 0

HIV/AIDS 0

GDP 0

Population 0

thinness 1-19 years 0

thinness 5-9 years 0

Income composition of resources 0

Schooling 0

dtype: int64

# Save the cleaned dataset

dataset.to\_csv('cleaned\_data.csv', index=False)

print("Cleaned dataset saved as 'cleaned\_data.csv'.")

Cleaned dataset saved as 'cleaned\_data.csv'.