import pandas as pd

import numpy as np

import matplotlib.pyplot as plt

import seaborn as sns

# Load the Excel file

file\_path = r"C:\Users\sreev\OneDrive\Desktop\diabetes analysis\diabetes.xlsx"

df = pd.read\_excel(file\_path)

columns\_with\_zeros = ['Glucose', 'BloodPressure', 'SkinThickness', 'Insulin', 'BMI']

# Replace zeros with NaN for specific columns

df[columns\_with\_zeros] = df[columns\_with\_zeros].replace(0, np.nan)

df[columns\_with\_zeros] = df[columns\_with\_zeros].fillna(df[columns\_with\_zeros].median())

# Display basic statistics

print("Basic statistics:")

print(df.describe())

# Plot histograms

df.hist(figsize=(12, 10))

plt.suptitle('Histograms of numerical features')

plt.show()

# Plot box plots

plt.figure(figsize=(12, 10))

df.boxplot(column=columns\_with\_zeros)

plt.title('Box plots of features with previously zero values')

plt.show()

# Plot a pairplot

sns.pairplot(df, hue='Outcome', diag\_kind='kde')

plt.suptitle('Pair plots of features colored by Outcome', y=1.02)

plt.show()

# Correlation heatmap

plt.figure(figsize=(10, 8))

correlation\_matrix = df.corr()

sns.heatmap(correlation\_matrix, annot=True, cmap='coolwarm', vmin=-1, vmax=1)

plt.title('Correlation heatmap')

plt.show()

# Save the cleaned data to a new Excel file

output\_path = './diabetes\_cleaned.xlsx'

df.to\_excel(output\_path, index=False)

print("Cleaned data saved to:", output\_path)







